UNIVERSIDAD NACIONAL DEL NORDESTE

FACULTAD DE CIENCIAS AGRARIAS

TRABAJO DE TESIS PRESENTADO POR:

Licenciada en Química Roxana Noelia Villafañe

TITULO:

"Estudio de la composición mineral de forrajeras nativas de la provincia de Corrientes. Propuesta de modelos quimiométricos para evaluar propiedades químicas y eventual origen geográfico"

(VERSIÓN CORREGIDA Y AMPLIADA)

PARA OPTAR AL GRADO ACADEMICO DE

Doctor de la Universidad Nacional del Nordeste en Recursos Naturales

Director: Dr. Roberto G. Pellerano

Co-Directora: Dra. Silvia M. Mazza

-2017-

A mis Padres

Por enseñarme que siempre

se puede empezar de nuevo

AGRADECIMIENTOS

A la Facultad de Ciencias Agrarias de la UNNE por permitirme completar mi formación académica.

Al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) por la beca otorgada que hizo posible la realización de este trabajo de tesis.

Al Dr. Roberto Pellerano por guiarme y ayudarme a lo largo de estos años de crecimiento académico.

A la Dra. Silvia Mazza por brindarme en la cátedra de Cálculo Estadístico y Biometría un lugar de trabajo.

A Eduardo Marchevsky por su guía y apoyo a lo largo de estos años.

De la cátedra de Cálculo Estadístico y Biometría quiero agradecer especialmente a Laura Gimenez, por haberme ayudado a dar los primeros pasos en la estadística en mi primer año de beca, y por ser alguien con quien siempre pude compartir pensamientos e ideas comunes.

A Griselda Bóbeda por ser una persona con la cual he podido compartir mi pensamiento y perspectiva libremente. Gracias por escucharme en los momentos más duros.

A Meli Hidalgo por ayudarme de muchas y diferentes maneras a lo largo de estos años.

A mis Padres, porque han sido de gran apoyo todos estos años, por entenderme y apoyarme SIEMPRE aunque seamos muy diferentes.

A mis amigos, por escucharme y apoyarme siempre, y por los buenos momentos compartidos.

Finalmente, a uno de los jurados, Dr. Carlos Acuña por las correcciones sugeridas, que han significado un aporte a esta tesis.

Nimm Dir Zeit um zu träumen,

es ist der Weg zu den Sternen.

(Tómate el tiempo para soñar,

este es el camino hacia las estrellas)

RESUMEN

Las Indicaciones Geográficas y Denominaciones de Origen constituyen herramientas que permiten diferenciar y hacer distinguible la calidad de un producto relacionada con su origen geográfico. Asimismo, otorgan protección legal al nombre, al producto, al grupo interesado, y tienden a promover el desarrollo rural. La búsqueda de sistemas que permitan autenticar el origen de un alimento de manera objetiva, mediante el estudio de su composición química ha sido el objeto de esta tesis. En este contexto, las variaciones en la abundancia isotópica natural y variaciones en las concentraciones de elementos minerales demuestran un excelente desempeño como "trazadores" de origen geográfico de los alimentos.

En los últimos cinco años el número de artículos científicos relacionados con la determinación del origen geográfico de alimentos ha sufrido un continuo aumento. Sin embargo, la producción científica relacionada con sistemas de trazabilidad químico de alimentos producidos en la región Nordeste Argentina resulta aún escasa. El objetivo general de esta tesis ha sido estudiar la presencia de posibles marcadores químicos de identidad en las partes aéreas de plantas utilizadas como forraje de la Provincia de Corrientes. Con este objetivo se analizaron las concentraciones de 18 elementos a nivel de vestigios por Espectrometría de Masas con Plasma Acoplado Inductivamente (ICP-MS), que es una técnica de análisis inorgánico elemental e isotópico capaz de determinar y cuantificar la mayoría de los elementos de la tabla periódica en un amplio rango dinámico. Las especies vegetales estudiadas fueron: *Desmodium incanum*, *Schizachyrium microstachyum* y *Andropogon lateralis,* seleccionadas de acuerdo a su disponibilidad y frecuencia de uso para la producción pecuaria de la zona de estudio.

Se determinaron las influencias de factores tales como, la serie de suelo donde fueron recolectadas las diferentes muestras y la relación con el sitio geográfico de muestreo. Todos los resultados fueron analizados mediante algoritmos estadísticos y de aprendizaje automático que permitieron modelar las variaciones presentes en los mismos y proponer modelos predictivos de origen geográfico para muestras problema futuras desconocidas.

En primera instancia se realizó un análisis exploratorio de datos mediante análisis de componentes principales (PCA) y un análisis de conglomerados (HCA), obteniéndose resultados que indicaron la posibilidad de proponer modelos para agrupar las muestras de acuerdo a su origen geográfico. Al utilizar como criterio de clasificación a la serie de suelo, se obtuvieron valores de exactitud superiores al 90% para las tres especies utilizando técnicas de clasificación multivariante relativamente sencillas de naturaleza paramétrica. Por otro lado, al considerar el sitio geográfico de obtención de las muestras, se hizo necesario utilizar técnicas de elevada complejidad, tales como máquinas de vectores soporte (SVM) o el algoritmo de árboles aleatorios (RF), para lograr resultados satisfactorios.

Los principales aportes de esta tesis se traducen en una profundización del conocimiento de la composición química mineral de plantas forrajeras nativas de la Provincia de Corrientes, abarcando no solo la concentración de elementos de interés nutricional, sino también elementos a nivel de vestigios sin función fisiológica demostrada o conocida; así también se logró la propuesta de modelos quimiométricos que permiten establecer con seguridad la procedencia geográfica de forrajes utilizados en la producción pecuaria.

ABSTRACT

The protected designation of origin (PDO) and protected geographical indication (PGI) are tools that allow to differentiate and distinguish the quality's product related to its geographical origin. Also, they provide legal protection to the name and the product, and tend to promote rural development. Searching these systems that enable to authenticate a product's origin through the study of the mineral chemistry is the aim of this thesis. In this context, variations in the natural abundance and variations in mineral concentrations demonstrate an excellent performance as tracers of geographical origin of foods.

In the last five years the number or publications related to geographical origin of food has been increased. Although, the scientific production related to chemical traceability systems produced in the northeast of Argentina is scarce. The general objective of this thesis has been to study the presence of possible tracers in aerial parts of plants that are used as forage in Corrientes Province. The concentrations of 18 elements were determined by ICP-MS that is capable to determine and quantify the majority of elements in the periodic table in a wide dynamic range. The botanic species studied were: *Desmodiun incanum*, *Schizachyrium microstachyum* and *Andropogon lateralis*, that were selected according to its availability and frequent usage in the livestock production in the study zone. Environmental factors, such as, soil series where have been collected and the relationship with geographical origin were studied. Every result was analyzed by statistical and machine learning algorithms that allowed us to model the mineral data and propose predictive models of geographical origin for unknown samples in the future.

In the first place, an exploratory data analysis was performed, and the results showed that a predictive analysis would be able to perform in order to cluster samples according to its geographical origin. When soil series were used as classificatory criteria, 90% of accuracy values were obtained using models relatively easy to implement. Considering the geographical origin another more complex techniques were necessary to implement, such as, support vector machines and random forest.

The main contribution of this work is the mineral composition data of native forages from Corrientes, Argentina; and not only the nutritional mineral concentrations but also the trace element concentrations with unknown physiological function. Also, it is a valuable contribution the chemometric data modeling that allows establishing the geographical origin of forages used for livestock.

Producción científica

Durante el desarrollo de la presente tesis de doctorado ha dado origen a los siguientes trabajos

PUBLICACIONES EN REVISTA CON REFERATO

- Toxic Trace Element Contents in Gluten-free Cereal Bars Marketed in Argentina publicado en *International Journal of Celiac Disease* 3 (1) 2015 **ISSN** 2334 - 3486 doi: 10.12691/ijcd-3-1-4 Hidalgo, MJ; **Villafañe, RN**; Sgroppo, SC; Marchevsky, EJ; Pellerano, RG.
- Tracing the geographical origin of argentinean lemon juices based on trace element profiles using advanced chemometric techniques. Microchemical Journal 129 (2016) 243-248 (Elsevier) doi: 10.1016/j.microc.2016.07.002 Gaiad, JE; Hidalgo, MJ; **Villafañe, RN**; Marchevsky, EJ; Pellerano, RG.
- Non-essential element concentrations in brown grain rice: assessment by advanced data mining techniques. Environmental Science and Pollution Research – (Springer) doi: 10.1007/s11356-017-9017-2 **Villafañe, RN**; Hidalgo, MJ; Piccoli, AB; Marchevsky, EJ; Pellerano, RG.
- Perfil mineral en los pastizales de *Andropogon lateralis* y *Sorghastrum setosum* (Gramineae) en Corrientes, Argentina. – Revista de la Facultad de Ciencias Agrarias (UNCuyo) Aceptado. En Prensa. Bernardis, AC; **Villafañe, RN**; Pellerano, RG; Marchevsky EJ Disponible en: http://revista.fca.uncu.edu.ar/images/stories/pdfs/En_prensa/Bernardis.pdf

PUBLICACIONES EN CONGRESOS

2013

- Assessment of mineral content in aerial parts of *Schizachyrium microstachyum* (Poaceae). XXXI Reunión científica anual de la Sociedad de Biología de Cuyo. Mendoza, Argentina. Noviembre 2013. Organiza: Cuyo Biology Society
- Non-essential trace element uptake in soybean seeds. XXXI Reunión científica anual de la Sociedad de Biología de Cuyo Mendoza, Argentina. Noviembre 2013. Organiza: Cuyo Biology Society
- Caracterización de dos especies forrajeras nativas de la provincia de Corrientes utilizando métodos quimiométricos basados en su composición mineral. 7mo Congreso Argentino de Química Analítica. Mendoza, Argentina. Octubre 2013. Organiza: Asociación de Químicos Analíticos.

2014

- Análisis Exploratorio del contenido multielemental de partes aéreas de *Desmodium incanum* (DC) provenientes de la provincia de Corrientes. XXX

Congreso Argentino de Química. Buenos Aires, Argentina. Octubre 2014. Organiza: Asociación Química Argentina.

- Determinación de bario, estroncio y litio en hierbas medicinales y sus infusiones consumidas en la región norte de Argentina. VII Congreso Iberoamericano de Física y Química Ambiental. Valparaíso, Chile. Octubre 2014. Organiza: Sociedad iberoamericana de Física y Química Ambiental.

2015

- Caracterización multivariada de forrajes nativos de Corrientes y rizosfera de acuerdo a su contenido mineral. 8vo Congreso Argentino de Química Analítica. La Plata, Buenos Aires, Argentina. Noviembre 2015. Organiza: Asociación de Químicos Analíticos.
- Determinación multielemental y clasificación de limones argentinos de acuerdo a su origen geográfico. 8vo Congreso Argentino de Química Analítica. La Plata, Buenos Aires, Argentina. Noviembre 2015. Organiza: Asociación de Químicos Analíticos.

2016

- Elementos traza no-esenciales y tóxicos para clasificación geográfica de muestras de arroz del Nordeste argentino. En el 5to Simposio Internacional de Biotecnología e Ingeniería Ambiental. Organiza: Instituto de Investigación e Ingeniería Ambiental (3iA), UNSAM. Buenos Aires, Argentina. Julio 2016.
- Clasificación Quimiométrica de muestras de Andropogon lateralis (Nees) proveniente de la provincia de Corrientes. En el XXXI Congreso Argentino de Química. Organiza: Asociación Química Argentina. Capital Federal, Buenos Aires. Octubre 2016.

CURSOS REALIZADOS

- Metodología de la Investigación Científica. FCA UNNE Carga Horaria: 60 hs
- Bioestadística y Diseño Experimental FCA UNNE Carga Horaria: 80 hs
- Modelos de Regresión Lineal y No Lineal. FCA UNNE. Carga Horaria: 30 hs
- Desarrollo y optimización de métodos espectrométricos aplicados a muestras complejas. FaCENA UNNE Carga Horaria: 60 hs
- Métodos de Análisis Multivariante. FCA UNNE Carga Horaria: 80 hs
- Diseño Experimental y Optimización de sistemas con múltiples respuestas. FaCENA UNNE Carga Horaria: 32 hs.
- Análisis Multivariado. FCA UNC Carga Horaria: 40 hs
- Espectroscopia de Plasma por Láser y sus Aplicaciones. FCQ UNC. Carga Horaria: 30 hs.
- Escritura de artículos científicos y tesis. Secretaría General de Ciencia y Técnica. Carga Horaria: 45 hs.
- Programación en R. FCEFyN – UNC. Carga Horaria: 40 hs.

- Análisis de datos multivariados para la calibración en Química Analítica. FBCB - UNL. Carga Horaria: 45 hs.
- Cursado de los trabajos prácticos de la materia Morfología de Plantas Vasculares. 89% de aprobados.

Horas totales cursadas: 542 hs. de un total de 11 cursos realizados y el cursado de los laboratorios de una materia de grado.

CURSOS ONLINE REALIZADOS

- Machine Learning Specialization: Classification – University of Washington (Coursera plataforma online)
- Machine Learning Specialization: Regression - University of Washington (Coursera plataforma online)
- Machine Learning Specialization: Foundations - University of Washington (Coursera plataforma online)
- Python for Data Science and Machine Learning (Udemy plataforma online)
- Ensemble Machine Learning: Random Forest and Adaboost (Udemy plataforma online)
- Learning from data – California Institute of Technology – (edX plataforma online)
- Statistical Learning – Stanford University – (Stanford Lagunita)
- Advanced Data Mining with WEKA - University of Waikato (plataforma online)
- More Data Mining with WEKA – University of Waikato (plataforma online)
- Data Mining with WEKA - University of Waikato (plataforma online)

ABREVIATURAS DE LA TESIS

ACC ..... Exactitud

AUC ..... Área bajo la curva

CHAV ..... Chavarría

CIC .....Capacidad de Intercambio iónico

ETAAS .....Espectrometría de absorción atómica electrotérmica

FAAS ..... Espectrometría de absorción atómica

ICP-MS .....Espectroscopía de masas por plasma acoplado inductivamente

HCA.....Análisis de conglomerados

IGP .....Productos de indicación geográfica protegida

LDA ..... Análisis Discriminante Lineal

m/z ..... Relación carga/masa

ntree..... Número de árboles

PCA ..... Análisis de componentes principales

PDO .....Productos designados de origen

PF ..... Paso Florentín

PMP ..... Pampín

PN ..... Paso Naranjito

PRE ..... Precisión

RMN .....Resonancia magnética nuclear

RF ..... Random Forest

RP ..... Ramada Paso

SENS/REC ..... Sensibilidad

SC ..... San Cosme

SM ..... San Miguel

SPEC ..... Especificidad

SVM ..... Support vector machines

UV-Vis ..... UV-Visible

INDICE DE TABLAS

INDICE DE FIGURAS

Tabla de contenido

SECCIÓN I: INTRODUCCIÓN

1) Capítulo I

1.1 Introducción

El interés de los consumidores de conocer el origen geográfico de los productos agropecuarios y en especial el de los alimentos ha crecido y adquirido una importancia trascendental en muchos países del mundo. Históricamente el consumo de ciertos productos se ha visto asociado a un determinado punto geográfico. El origen geográfico puede ser un elemento esencial para la autenticidad de un determinado producto alimentario, para proteger productos regionales y para confirmar las características de calidad relacionadas con el lugar de origen del alimento (Drivelos, S. A. y Georgiou, C. A. 2012, Luykx, D. M. A. M. y van Ruth, S. M. 2008).

Para esto hay que disponer de un sistema de "trazabilidad" que significa la habilidad para rastrear y seguir un alimento de origen animal o la sustancia que pretenda serlo, o que se espera sea incorporado a un alimento o pienso, a través de todas las etapas de la producción, procesado y distribución (Bertacchini, L. et al. 2013, Bosona, T. y Gebresenbet, G. 2013). Las técnicas de "huella dactilar" describen a una variedad de métodos analíticos que pueden medir la composición de productos alimentarios de una manera no-selectiva, esto es, mediante la colección de un espectro y/o determinación de la composición multielemental. El procesado matemático de la información contenida en esas "huellas dactilares" permitiría la caracterización del producto alimentario. Los métodos que pueden proporcionar un perfil mineral característico se pueden usar para la determinación del origen geográfico y, por lo tanto, proporcionar una herramienta valiosa para la autenticación y trazabilidad de alimentos (Esslinger, S. et al. 2014).

Los elementos minerales juegan un papel importante en la trazabilidad geográfica de los alimentos, así la comprensión de los perfiles químicos y relaciones entre la disponibilidad en los distintos tipos de suelo y presencia en los tejidos ayuda a determinar la correspondencia entre la acumulación y sus propiedades particulares, muchas veces relacionadas con factores que le otorgan valor agregado al poder determinar y cuantificar su procedencia geográfica. Resulta importante destacar que la composición inorgánica de los vegetales se ve afectada por factores tales como la naturaleza química y tipo de suelo, acidez del suelo, disponibilidad de componentes inorgánicos e incluso las condiciones climáticas como la humedad y la temperatura (Ariyama, K. y Yasui, A. 2006). Es decir que la composición mineral de los vegetales se verá fuertemente influenciada por las condiciones en las que fueron producidas (Watson, C. A. et al. 2012).

De la revisión de la literatura disponible surge que en la actualidad se ha utilizado la composición mineral para determinar la procedencia geográfica de diversos alimentos, entre ellos se puede destacar: determinación de la procedencia geográfica de vino (Dutra, S. V. et al. 2013, Geana, I. et al. 2013, Šelih, V. S. et al. 2014), mieles (Batista, B. L. et al. 2012, Di Bella, G. et al. 2015), arroz (Cheajesadagul, P. et al. 2013, Chung, I.-M. et al. 2015, Maione, C. et al. 2016), leche (Magdas, D.-A. et al. 2016, Nečemer, M. et al. 2016), entre otros.

1.2 Perfil mineral de los vegetales como marcadores de autenticidad

La determinación de autenticidad de los alimentos es un tema importante en el control y la seguridad de los alimentos de calidad. El brote de enfermedades transmitidas por los alimentos en todo el mundo ha aumentado la conciencia de los

consumidores sobre la calidad y seguridad de los alimentos, los aspectos relacionados con su origen geográfico y las prácticas agrícolas en la producción de alimentos han asumido una gran importancia. La calidad de un producto es un problema para todos los agricultores y el comprador, ya se trate de mercancías producidas con las normas básicas o con productos de alta calidad. Los agricultores de todo el mundo deben ofrecer productos de alta calidad para mantener la competitividad y la rentabilidad. Tendiendo a lograr esto, la legislación establece unos requisitos estrictos a fin de garantizarles las normas de todos los productos.

Las técnicas analíticas, tales como la cromatografía líquida de alta resolución, cromatografía de gases, espectroscopía UV-Vis, la espectroscopia de infrarrojo, resonancia magnética nuclear (RMN), han sido empleadas para controlar la presencia de los principales componentes de la muestra o algunos compuestos orgánicos que pueden ser característica de un producto de origen específico. Sin embargo, el rango normal de compuestos orgánicos en los alimentos varía con algunos parámetros agrícolas como la fertilización, las condiciones climáticas en el año de cultivo, la historia de los campos y la variedad o especie, así como la ubicación y características del suelo, por lo que a veces es difícil ser contundente acerca de la autenticidad a partir de la determinación de los componentes orgánicos. Como resultado, hay una demanda continua de técnicas eficaces para el control de la autenticidad en productos alimenticios. El contenido de minerales es una excelente alternativa debido a la correlación entre elementos a nivel de vestigios, el tipo de suelo y las condiciones de cultivo ambientales.

Debido a esto, la evaluación de los contenidos de elementos a nivel de vestigios ha sido propuesta para determinar el origen geográfico de las muestras. Por lo tanto,

la determinación de los perfiles de minerales de productos designados de origen (PDO) y productos de indicación geográfica protegida (IGP), es una tarea importante por razones de salud y de autenticación y, por lo tanto, las técnicas de un solo elemento (por ejemplo, espectrometría de absorción atómica -FAAS- y la espectrometría de absorción atómica electrotérmica –ETAAS-), así como técnicas multi-elementales (por ejemplo, plasma acoplado inductivamente espectrometría de emisión óptica -ICP- OES- y acoplado inductivamente espectrometría de masas con plasma -ICP-MS-) se han empleado con éxito en la autenticación de productos designados de origen (Gonzálvez, A. y de la Guardia, M. 2013).

Diferentes pasos se siguen para la determinación mineral de un producto determinado. El problema analítico debe ser bien definido: una clara identificación del problema, el contexto del problema, el tipo de información necesaria. Un segundo paso incluye la selección de un procedimiento analítico para la muestra bajo estudio. La técnica de análisis químico debe incluir no sólo el tipo de equipo sino también el pre-tratamiento de las muestras. En la Fig 1.1 se observa las técnicas instrumentales utilizadas para determinación elemental en diferentes productos.

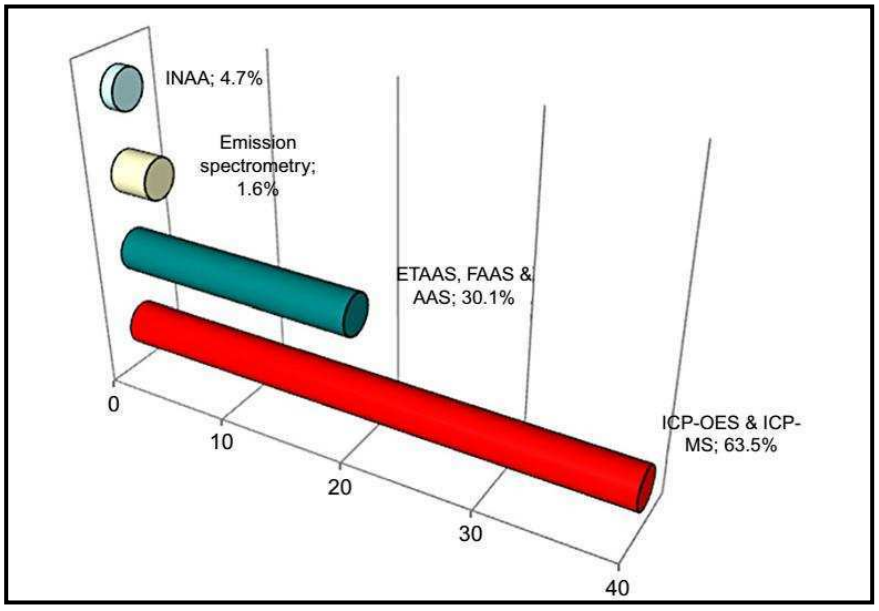


Fig. 1.1 Gráfico de barras con la distribución de técnicas analíticas utilizadas para determinar contenido mineral en diferentes productos

1.3 El suelo

El suelo es uno de los elementos clave en todos los ecosistemas terrestres en los que cumple cinco funciones esenciales:

1) es el soporte físico para el crecimiento de las plantas, les proporciona el agua, el aire y los nutrientes esenciales para su desarrollo y las protege de ciertos elementos fitotóxicos, como así también de cambios bruscos de temperatura. Las propiedades del suelo son las que determinan el tipo de vegetación de un ecosistema y la producción de biomasa,
2) es el factor principal que controla el flujo del agua en el ciclo hidrológico, así como de las especies químicas dentro de los ciclos biogeoquímicos. Aunque es una capa muy delgada, su composición y propiedades pueden ser muy diferentes de las del substrato geológico y, en su ausencia, los ciclos hidrológicos y biogeoquímicos serían completamente distintos,
3) actúa como sistema de reciclado natural donde los productos de desecho son transformados en otros compuestos disponibles para las generaciones siguientes, o bien, son retenidos actuando como sumidero, por ejemplo, CO_2
4) es el hábitat de una gran variedad de organismos vivos, desde microorganismos hasta reptiles. Por último,
5) es el soporte para la infraestructura, necesario para el desarrollo de la vida humana.

Resumiendo, es la interfase donde se producen los procesos de interacción entre la litosfera, atmósfera, hidrosfera y biosfera. Las definiciones de suelos son muy variadas dependiendo de su función. Desde el punto de vista agrícola, se define como una

entidad natural dinámica de la superficie de la tierra donde crecen las plantas, compuesto de minerales, materiales orgánicos y organismos vivos (Brady, N. C. y Weil, R. R. 2002)

El suelo es un medio heterogéneo muy complejo que consta de una fase sólida (la matriz del suelo) constituida por minerales, materia orgánica y una fase fluida constituida por la solución y el aire del suelo, que interaccionan entre ellas y con los elementos que entran en el sistema del suelo (Alloway, 1995). La proporción relativa de estos cuatro componentes le confiere sus propiedades, su comportamiento y su productividad. En general, el 45% del suelo es el componente mineral, el 5% la fracción orgánica, un 25% lo constituye el aire y el otro 25% el agua alojada en los poros. La fase acuosa del suelo, denominada solución del suelo es en su mayoría agua que contiene especies químicas procedentes de los procesos químicos y bioquímicos del suelo y del intercambio con la hidrosfera y biosfera. Además, contiene oxígeno y dióxido de carbono disueltos. El suelo es el medio que proporciona agua a las plantas y es el vehículo de transporte e intercambio de los nutrientes entre las partículas del suelo y las raíces de las plantas. Normalmente, debido al pequeño tamaño de las partículas del suelo y la presencia de poros, la fase acuosa no es totalmente independiente de la fase sólida. La mayoría de los elementos minerales se encuentran presentes como iones, H^+, Ca^{2+}, Mg^{2+}, Na^+ y K^+ y pequeñas cantidades de Fe^{2+}, Mn^{2+} y Al^{3+}. Estos tres últimos cationes se suelen encontrar parcialmente hidrolizados o en forma de complejos con ligandos orgánicos. Los aniones mayoritarios son HCO_3^- , CO_3^{2-}, SO_4^{2-} y Cl^- que pueden estar al mismo tiempo unidos a H^+ o a otros iones como por ejemplo el Ca^{2+}. La absorción de elementos traza por parte del suelo ha sido y es un tema de vital importancia no sólo para la agricultura sino también para temas

ambientales como por ejemplo la remediación y la descontaminación de suelos (Antoniadis, V. et al. 2017).

Las series de suelo son Unidades taxonómicas, son suelos semejantes, desarrollados a partir de un mismo material originario, con igual secuencia de horizontes y demás características morfológicas principales similares. Las unidades taxonómicas, se utilizan para clasificar a los suelos dentro de un sistema, pero no indican cómo están distribuidos en el terreno. Para ello se recurre al uso de las unidades cartográficas, que son delineaciones continuas en el plano que indican el agrupamiento de suelos en asociaciones o complejos. Para este estudio las muestras fueron tomadas de las Serie de suelo Chavarría y Pampín (Tabla 1.1).

Tabla 1.1 Clasificación taxonómica de los suelos empleados en este trabajo

Orden	Suborden	Gran Grupo	Sub grupo	Familia	Serie
ENTISOLES	ACUENTES	PSAMACUENTES	SPODICOS	Arenosa	Chavarría
			TIPICOS	Arenosa	Pampín

1.3.1 Series de suelo

1.3.1.1 Serie Chavarría

Constituye una de las series de suelo de mayor distribución y superficie dentro de la provincia de Corrientes. Se ubica en relieve normal, en posición de media loma a media loma baja, con pendientes de 1 a 1.5%, en planicies arenosas pardo amarillentas. El tapiz vegetal está compuesto por pajonales de *Andropogon lateralis*, acompañado de *Axonopus sp*, *Schizachirium sp*., *Sporobolus sp*. y otros de hábitos húmedos como *Ciperáceas* y *Centella*. Son suelos de muy baja fertilidad, con escaso tenor de materia orgánica, bajo contenido de bases de cambio y capacidad de intercambio catiónico (CIC), débilmente ácidos y de pobre retención de humedad en los horizontes superiores. Los suelos presentan muy severas limitaciones que

restringen la elección de plantas y requieren un manejo cuidadoso. Las principales limitantes se refieren al exceso de humedad con sobresaturación por tiempos prolongados, además de su baja fertilidad natural. El uso actual es la ganadería extensiva, no obstante, es utilizado para forestación y agricultura, con los consiguientes riesgos, si no se mejoran las condiciones de drenaje y fertilidad. Se ubica en la Clase IVw y el Indice de Productividad es de 16 (Escobar, et al., 1996).

1.3.1.2 Serie Pampín

Se ubica en relieve normal, posición de loma, con pendientes de 1 a 1,5%. El tapiz vegetal está compuesto de *Paspalum notatum, Cynodon sp*., *Sporobolus sp*. y *Axonopus sp*. El escurrimiento es medio, la permeabilidad moderadamente lenta y el drenaje moderado a imperfecto, con peligro de sobresaturación con agua en épocas de grandes lluvias. Son suelos profundos, compuestos por un manto arenoso de 120 cm de espesor, en donde se diferencia un horizonte superficial ócrico, enriquecido por materia orgánica, arenoso, pardo a pardo grisáceo oscuro, fuertemente ácido. La profundidad efectiva es mayor a 100 cm, y la retención de humedad es muy baja. Presenta muy severas limitaciones que restringen la elección de cultivos por su baja fertilidad, susceptibilidad a la erosión eólica y exceso de humedad en períodos de lluvias excesivas. Generalmente se utilizan para cultivos anuales en forma ocasional. El uso actual es el de campo natural de pastoreo, forestación y cultivos de hortalizas. La Clase por Capacidad de Uso es IVw y el Índice de Productividad es 20 (Escobar, et al., 1996).

1.4 Especies vegetales

En este trabajo de tesis se estudiaron tres especies vegetales:

1.4.1 *Andropogon lateralis*

Nombre común: Paja colorada. Es una planta de matas altas y porte erecto. Produce abundantes cañas florales de hasta 150 cm de altura, cuando maduran son de color pardo-rojizo por esta característica recibe el nombre de Paja colorada. Es una especie perenne de ciclo estival. Está presente en casi todas las regiones, constituyendo la gramínea que domina los pastizales de la provincia en las lomadas arenosas, campos altos con afloramientos rocosos, malezales (bajos inundables temporariamente) (Sanpedro, 2002). Al ser una especie forrajera dominante, su cobertura está alrededor del 60 %, en otras es un componente secundario (Trindade 1999; Nabinger 2009).

1.4.2 *Schizachyrium microstachyum*

Pertenece a la tribu Andropogoneae, es de distribución pantropical. Habitan campos abiertos, primitivos y sabanas tropicales, desapareciendo donde se hace agricultura. Pueden ser consumidas por el ganado en épocas de escasez de forrajes, antes del encañado.

1.4.3 *Desmodium incanum*

Nombre común: “Pega-Pega”. Especie rastrera a erecta de 15-50 cm de altura, perenne de ciclo primavero-estival. Alcanza su mayor crecimiento en febrero-abril. Dentro de las leguminosas nativas es la especie más difundida en el Centro-sur de Corrientes, produciendo forraje moderadamente tierno, que es bien consumido por los animales.

1.5 Elementos a nivel de vestigios

La nutrición vegetal se relaciona con el abastecimiento y absorción de compuestos químicos necesarios para el crecimiento y metabolismo de plantas. Los compuestos requeridos por los vegetales se denominan nutrientes. El 90-95% del peso

seco del material vegetal está constituido por C, O e H, que son los principales constituyentes de los compuestos orgánicos, y el 5-10% restante corresponde a otros elementos cuya presencia es esencial para completar su desarrollo normal y su ciclo biológico. Debido a su papel fisiológico se les llama elementos esenciales y, de acuerdo con la concentración en que son requeridos por la planta, se clasifican en macro y micronutrientes.

1.5.1 Clasificación de nutrientes

Los elementos esenciales se definen como aquellos sin los cuales las plantas no pueden completar su ciclo de vida, son irremplazables por otros elementos, y están involucrados directamente en el metabolismo de la planta (Adriano, D. C. 2001, Fageria, N. K. et al. 2002, Kabata-Pendias, A. 2010). Basados en la cantidad necesaria, los nutrientes se dividen en macronutrientes o micronutrientes (elementos a nivel de vestigios). Se los llama elementos minoritarios indicando que su concentración es mínima con relación a los mayoritarios. Los elementos a nivel de vestigios se definen como elementos que están presentes a bajas concentraciones (mg/kg o menor) en la mayoría de suelos, plantas, y demás organismos vivos (Adriano, D. C. 2001). El Cu, Zn, Fe, Mn, Mo, y B son esenciales para el crecimiento normal de las plantas, Cu, Zn, Fe, Mn, Mo, Co y Se son esenciales para el crecimiento y la salud de animales y seres humanos, y Cu, Zn, Pb y Cd son los más ambientalmente importantes que han sido reportados que causan la contaminación del suelo, el agua y la cadena alimentaria (He, Z. L. et al. 2005). La acumulación de estos nutrientes en plantas generalmente sigue el siguiente patrón: Mn > Fe > Zn > B > Cu > Mo. Este orden puede cambiar según las especies de plantas y las condiciones de crecimiento. A pesar de que los

micronutrientes se requieren en pequeñas cantidades por las plantas, su influencia es tan importante como los macronutrientes.

Esta clasificación tiene una validez relativa, ya que, en algunos casos, ciertos macronutrientes se acumulan en cantidades menores que otros micronutrientes. Por otro lado, además de los nutrientes se pueden encontrar en la planta otros elementos, sin función biológica conocida hasta ahora, y otros cuya presencia en la planta conduce a disfunciones, ya que tienden a acumularse y son altamente tóxicos (Cd, Hg, Pb, Sn y Bi).

Algunos son nutrientes y esenciales para las plantas y animales, micronutrientes (como el Mn, Mo, Cu, Co, Zn, Sc, y V); mientras que otros elementos (como el Ni, Sn, y Cr) son esenciales solamente para los animales. No obstante, cuando estos elementos están presentes en sistemas ambientales a concentraciones superiores a ciertos niveles, bien sea a causa de desequilibrios naturales o sobre todo debido a su introducción antropogénica, pueden ser tóxicos para los seres vivos (Doménech X. y Peral J., 2006).

La consideración sobre la esencialidad de un determinado elemento varía en el tiempo de acuerdo a estudios que permitan determinarla. En esta tesis se trabajaron con los siguientes elementos esenciales: Co, Cu, Cr, Mo, Se y Zn, probablemente esenciales: Al, B, Cd, Li, Ni, Sn, Sr, Ti, V; y con función incierta: Sb y Tl (Bernardis, A. et al. 2016, NRC 2001)

1.5.2 Fuentes de Elementos a nivel de vestigios

Los elementos entran a un agro-sistema a través de procesos naturales y antropogénicos. El suelo hereda elementos minoritarios desde la *roca madre*. La idea de roca madre hace mención a aquella roca que define la matriz mineral de un suelo.

Se han encontrado algunos suelos que tienen una gran cantidad de elementos a nivel de vestigios, los cuales son tóxicos para las plantas y la vida animal, debido a las altas concentraciones en la roca madre. Los procesos antropogénicos incluyen la entrada de elementos a nivel de vestigios a través del uso de fertilizantes, desechos orgánicos, desperdicios industriales y municipales, irrigación y depósitos húmedos o secos. Estos procesos contribuyen a valores diferentes de concentración en los elementos a nivel de vestigios en el agro-sistema.

Sólo unas pequeñas porciones de elementos a nivel de vestigios en el suelo están biodisponibles. La movilidad y la disponibilidad de los elementos a nivel de vestigios están controladas por muchos procesos químicos y bioquímicos como la precipitación-disolución, adsorción-desorción, reacciones de complejación-disociación y oxidación-reducción. No todos los procesos son igualmente importantes para cada elemento, pero todos los procesos están afectados por el pH del suelo y los procesos biológicos. Entonces, es necesario entender algunas reacciones importantes en suelos que controlan la liberación de ciertos elementos específicos en el suelo y en el ambiente para superar problemas relacionados a la deficiencia y la contaminación de estos elementos (Adriano, D. C. 2001). La Fig 1.2 muestra el ciclo geoquímico generalizado a nivel de vestigios (Adriano, D. C. 2001).

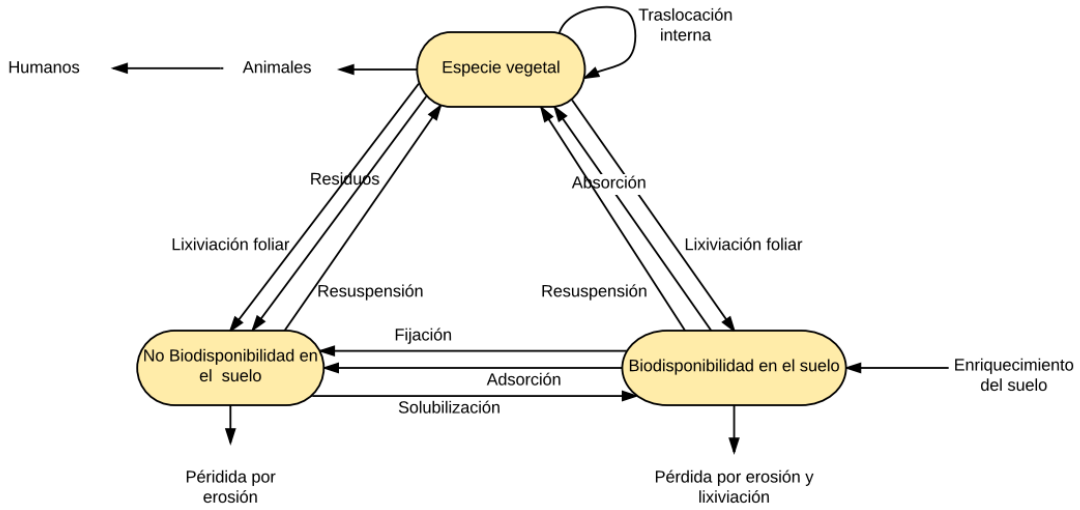


Fig. 1.2 Ciclo geoquímico generalizado de elementos a nivel de vestigios en un agro-sistema

El sistema planta-suelo actúa como una barrera efectiva en contra de elementos potencialmente tóxicos para animales (Ni, Cd) ya que el crecimiento de la planta se detendrá antes que estos elementos puedan ser absorbidos desde el suelo y acumulados en concentraciones que pueden ser dañinos para los animales (Adriano, D. C. 2001).

1.5.3 Biodisponibilidad de elementos a nivel de vestigios

Los elementos a nivel de vestigios del suelo se encuentran: 1) formando parte de los minerales primarios cuya meteorización puede tardar miles de años en producirse y en consecuencia, no son asimilables para las plantas; 2) formando parte de arcillas por sustituciones isomórficas del Fe y Al de las capas octaédricas; 3) ocluidos en óxidos de Fe y Mn; 4) formando complejos con la materia orgánica; 5) como cationes de cambio; y, finalmente, 6) como fases solubles presentes en la solución del suelo, en forma iónica o bien formando complejos, siendo estos fácilmente absorbidos por las plantas. La concentración total de elementos a nivel de vestigios en el suelo, por tanto, no refleja los niveles disponibles ya que solamente una parte pequeña se

encuentra en forma soluble y como consecuencia, disponible para las plantas (Lassat, 2001). Las formas solubles aparecen como resultado de reacciones químicas entre los materiales inorgánicos y orgánicos presentes en el suelo, y las fases acuosa y gaseosa (Fig 1.3).

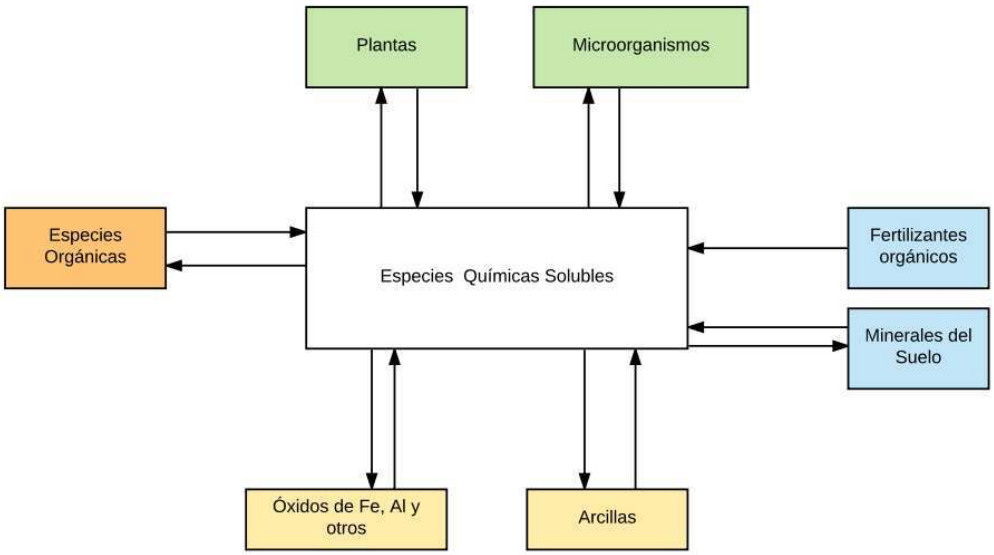


Fig. 1.3 Especies químicas solubles en el suelo

En definitiva, según la especie química en que se encuentre presente los elementos a nivel de vestigios en las diferentes fases del suelo, así será la disponibilidad relativa por las plantas y, por tanto, la incorporación en la biomasa. La fertilidad del suelo se debe corresponder con su capacidad de almacenar grandes cantidades de nutrientes cuando las necesidades de las plantas son más bajas o no existen, y disponer de grandes cantidades de nutrientes en solución mientras los necesitan las plantas. La biodisponiblidad se define como la fracción del metal que puede interactuar con un organismo biológico y ser incorporado a su estructura (Vangronsveld y Cunningham, 1998). La movilidad de minerales en el suelo es un fenómeno variable que depende de unas condiciones muy determinadas del suelo

que, en caso de alterarse, conducen a la liberación o retención de estos elementos. La capacidad de un suelo de suministrar nutrientes depende de las propiedades físicas, químicas y biológicas de éste y del clima. Las complejidades de estos procesos se pueden sintetizar teniendo en cuenta dos aspectos primordiales: las fases activas del suelo (arcilla y óxidos de Fe y Mn procedentes de la meteorización y materia orgánica) y las reacciones fisicoquímicas que se producen entre la fase sólida y la solución del suelo. Estas últimas son estrechamente dependientes del pH y del potencial redox del sistema, que a su vez dependen de la textura y estructura del suelo. Las fases activas proporcionan al suelo una de sus propiedades más importantes, la capacidad de intercambio catiónico (CIC).

1.5.4 Disponibilidad de elementos minerales en suelo

1.5.4.1 Capacidad de Intercambio Catiónico (CIC)

La capacidad de intercambio catiónico es función del contenido de arcilla y materia orgánica, fundamentalmente. En general cuanto mayor sea la capacidad de intercambio catiónico, mayor será la capacidad del suelo de fijar metales. El poder de adsorción de los distintos metales pesados depende de su estado de oxidación y del radio iónico hidratado; a mayor tamaño y menor estado de oxidación, menos fuertemente quedan retenidos. Dado que la carga neta de los materiales procede del equilibrio entre cargas permanentes y cargas variables, el CIC depende del pH (Bergaya, F. et al. 2011, Ulusoy, Y. et al. 2016).

1.5.4.2 pH

El pH es el factor más importante en el control de todas las reacciones fisicoquímicas implicadas en la disponibilidad de los elementos a nivel de vestigios. El

pH del suelo influye en la solubilidad, concentración en la solución del suelo, forma iónica, y la movilidad de micronutrientes en el suelo, y la consecuente absorción de estos elementos por las plantas. Como regla, la disponibilidad de B, Cu, Fe, Mn y Zn usualmente decrece, y el Mo incrementa cuando el pH aumenta. Estos nutrientes generalmente son siempre absorbidos en la superficie del suelo (Fageria, N. K., et al. 2002). La Tabla 1.2 resume importantes cambios en las concentraciones de micronutrientes influenciado por el pH del suelo y la consecuente absorción por plantas (Fageria, N. K., et al. 2002).

Tabla 1.2 Influencia del pH en la concentración de micronutrientes.

Micronutrientes	Influencia en la concentración
Cu	La solubilidad de Cu^{2+} es dependiente del pH y decrece 100 veces por cada unidad que se incrementa en el pH. La absorción también decrece.
Fe	La concentración de ion férrico (Fe^{3+}) y ferroso (Fe^{2+}) en la solución del suelo decrece 1000 veces 100 veces, respectivamente, por cada unidad que se incrementa el pH. En la mayoría de los suelos oxidados, la toma de hierro en las plantas decrece aumentando el pH.
Mn	La principal especie en la que se presenta el Mn es el Mn^{2+}, y las concentraciones decrecen 100 veces por cada unidad que aumenta el pH. En suelos extremadamente ácidos, la solubilidad del Mn^{2+} puede ser lo suficientemente alta para provocar problemas de toxicidad en las especies sensibles.
Mo	Arriba de un pH de 4.2, MoO_4^{-2} es mayoritario. Las concentraciones de estas especies se incrementan con el aumento de pH y la absorción en la planta también aumenta. La concentración de Mo se incrementa seis veces cuando el pH aumenta de 4,7 a 7,5.
Zn	La solubilidad del Zn es altamente dependiente del pH del suelo, y decrece 100 veces por cada unidad que se incrementa en el pH, y la absorción de las plantas disminuye como consecuencia.
Ni	El Ni^{2+} es relativamente estable en amplios rangos de pH de suelo y condiciones redox. De todas formas, la biodisponibilidad es usualmente más alto en suelos ácidos que en suelos alcalinos. A un pH 7 o superior, la retención y la precipitación se incrementa.
Co	La solubilidad y la biodisponibilidad de Co disminuye con pH extremos en el suelo.

La mayoría de los metales tienden a estar más disponibles a pH ácido, excepto As, Mo, Se y Cr, los cuales tienden a estar más disponibles a pH alcalino. En medios de pH moderadamente alcalino se produce la precipitación como hidróxidos. En medios muy alcalinos, pueden nuevamente pasar a la solución como hidroxicomplejos. Por otra parte, algunos metales pueden estar en la disolución del suelo como aniones solubles. Tal es el caso de los siguientes metales: Se, V, As y Cr. La asimilación de nutrientes del suelo por la planta está influenciada por el pH (Fageria, N. K., et al. 2002).

1.5.4.3 Potencial Redox (Condiciones de oxidación y reducción)

Las condiciones de oxidación-reducción (potencial redox) tienen efecto sobre aquellos iones que pueden tener varios estados de oxidación. Las condiciones redox pueden afectar indirectamente la movilidad de metales. Por ejemplo, en condiciones reductoras el Fe^{3+} se transforma en Fe^{2+} mucho más soluble. Así, muchos metales que están asociados o adsorbidos a hidróxidos de Fe y Mn son estables a potenciales redox bajos. Dependiendo de las condiciones químicas, se movilizan. También, en ambientes muy reductores el Fe se puede combinar con el S^{2-} hasta convertirse en pirita. Cuando los suelos y sedimentos contienen cantidades significativas de pirita y aumenta el potencial redox (creación de condiciones más oxidantes) el S^{2-} se oxida a SO_4^{2-} liberando cantidades de H_2SO_4, el suelo se acidifica fuertemente y los metales se hacen muy solubles (Fageria, N. K., et al. 2002).

1.6 Referencias bibliográficas

Adriano DC. (2001) Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals: Springer Science & Business Media.

Antoniadis V, Levizou E, Shaheen SM, Ok YS, Sebastian A, Baum C, Prasad MNV, Wenzel WW, Rinklebe J. (2017) Trace elements in the soil-plant interface: Phytoavailability, translocation, and phytoremediation–A review. Earth-Science Reviews.

Ariyama K, Yasui A. (2006) The Determination Technique of the Geographic Origin of Welsh Onions by Mineral Composition and Perspectives for the Future. Japan Agricultural Research Quarterly: JARQ.40:333-339.

Bernardis A, Villafañe R, Pellerano R, Marchevsky E. (2016) Perfil mineral en los pastizales de Andropogon lateralis y Sorghastrum setosum (Gramineae) en Corrientes, Argentina. Revista Facultad de Ciencias Agrarias UNCuyo.

Batista BL, da Silva LRS, Rocha BA, Rodrigues JL, Berretta-Silva AA, Bonates TO, Gomes VSD, Barbosa RM, Barbosa F. (2012) Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. Food Res Int.49:209-215.

Bergaya F, Theng BKG, Lagaly G. (2011) Handbook of Clay Science: Elsevier.

Bertacchini L, Cocchi M, Li Vigni M, Marchetti A, Salvatore E, Sighinolfi S, Silvestri M, Durante C. (2013) The Impact of Chemometrics on Food Traceability. Data Handling in Science and Technology.28:371-410.

Bosona T, Gebresenbet G. (2013) Food traceability as an integral part of logistics management in food and agricultural supply chain. Food Control.33:32-48.

Brady NC, Weil RR. (2002) The Nature and Properties of Soils: Prentice Hall.

Cheajesadagul P, Arnaudguilhem C, Shiowatana J, Siripinyanond A, Szpunar J. (2013) Discrimination of geographical origin of rice based on multi-element fingerprinting by high resolution inductively coupled plasma mass spectrometry. Food Chem.141:3504-3509.

Chung I-M, Kim J-K, Lee J-K, Kim S-H. (2015) Discrimination of geographical origin of rice (Oryza sativa L.) by multielement analysis using inductively coupled plasma atomic emission spectroscopy and multivariate analysis. Journal of Cereal Science.65:252-259.

Di Bella G, Lo Turco V, Potortì AG, Bua GD, Fede MR, Dugo G. (2015) Geographical discrimination of Italian honey by multi-element analysis with a chemometric approach. J Food Compos Anal.44:25-35.

Doménech X, Peral J. (2006) Química ambiental de sisemas terrestres: Reverté

Drivelos SA, Georgiou CA. (2012) Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. Trends in Anal Chem.40:38-51.

Dutra SV, Adami L, Marcon AR, Carnieli GJ, Roani CA, Spinelli FR, Leonardelli S, Vanderlinde R. (2013) Characterization of wines according the geographical origin by analysis of isotopes and minerals and the influence of harvest on the isotope values. Food Chem.141:2148-2153.

Escobar EH, Ligier HD, Melgar R, Matteio H, Vallejos O. (1996) Mapa de suelos de provincia de Corrientes 1:500.000 INTA:Centro Regional Corrientes.

Esslinger S, Riedl J, Fauhl-Hassek C. (2014) Potential and limitations of non-targeted fingerprinting for authentication of food in official control. Food Res Int.60:189-204.

Fageria NK, Baligar VC, Clark RB. (2002) Micronutrients in Crop Production. In: Adv Agron. Academic Press. p. 185-268.

Geana I, Iordache A, Ionete R, Marinescu A, Ranca A, Culea M. (2013) Geographical origin identification of Romanian wines by ICP-MS elemental analysis. Food Chem.138:1125-1134.

Gonzálvez A, de la Guardia M. 2013. Chapter 3 - Mineral Profile. In: Comprehensive Analytical Chemistry. Elsevier. p. 51-76.

He ZL, Yang XE, Stoffella PJ. (2005) Trace elements in agroecosystems and impacts on the environment. J Trace Elem Med Biol.19:125-140.

Kabata-Pendias A. (2010) Trace Elements in Soils and Plants, Fourth Edition: Taylor & Francis.

Lassat, M. (2001) The use of plants for removal of toxic metals from soils. US-EPA

Luykx DMAM, van Ruth SM. (2008) An overview of analytical methods for determining the geographical origin of food products. Food Chem.107:897-911.

Magdas D-A, Dehelean A, Feher I, Cristea G, Puscas R, Dan S-D, Cordea D-V. (2016) Discrimination markers for the geographical and species origin of raw milk within Romania. Int Dairy J.61:135-141.

Maione C, Batista BL, Campiglia AD, Barbosa Jr F, Barbosa RM. (2016) Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. Comput Electron Agric.121:101-107.

Nabinger C, de Faccio Carvalho PC (2009) Ecofisiología de sistemas Pastoriles: Aplicaciones para su sustentabilidad. Agrociencia 13:18-27

Nečemer M, Potočnik D, Ogrinc N. (2016) Discrimination between Slovenian cow, goat and sheep milk and cheese according to geographical origin using a combination of elemental content and stable isotope data. J Food Compos Anal.52:16-23.

Sanpedro, DH (2002) XIX Reunión del grupo técnico en forrajeras del Cono Sur zona campos. Mercedes, Corrientes, Argentina. Memorias. INTA EEA. Mercedes.

Šelih VS, Šala M, Drgan V. (2014) Multi-element analysis of wines by ICP-MS and ICP-OES and their classification according to geographical origin in Slovenia. Food Chem.153:414-423.

Trindade JPP, Cuadros LF, de e Rocha MG. (1999) Estimativa de taxa de crescimento de làminas foliares em afilhos de *Andropogon lateralis* Nees. Em pastagem natural submetida a manejos de quima e pastajo. In: Moraes. A. et al (ED.) Simposio Internal “Grassland Ecophysiology and Grazing Ecology”. Nais. Curitiba. 24 a 26/08/1999. Curitiba: UFPR. P. 280-283.

Ulusoy Y, Tekin Y, Tümsavaş Z, Mouazen AM. (2016) Prediction of soil cation exchange capacity using visible and near infrared spectroscopy. Biosyst Eng.152:79-93.

Vangronsveld J, Cunningham SD. (1998) Metal-Contaminated soils: In-situ inactivation and phytoremediation

Watson CA, Öborn I, Edwards AC, Dahlin AS, Eriksson J, Lindström BEM, Linse L, Owens K, Topp CFE, Walker RL. (2012) Using soil and plant properties and farm

management practices to improve the micronutrient composition of food and feed. J Geochem Explor.121:15-24.

Objetivos Propuestos

Objetivos Generales

El presente plan de trabajo se propone contribuir al conocimiento de la composición química mineral de especies nativas de plantas utilizadas como forrajes para la producción pecuaria en la provincia de Corrientes. Si bien los niveles de los elementos minerales en las muestras botánicas bajo estudio se ven afectados por numerosos factores de naturaleza diversa, es objeto de este plan también la determinación de los mismos en muestras de suelo de las regiones donde habitan estas especies, para explorar las influencias que pudieran ejercer sobre la composición global. Finalmente, el presente plan se propone desarrollar modelo quimiométricos que permitan conocer la presencia de trazadores químicos del origen geográfico de los productos naturales estudiados.

Objetivos Particulares

- Determinar composición química inorgánica de especies de plantas nativas forrajeras, mediante espectrometría de masas por plasma acoplado inductivamente (ICP-MS).
- Detectar la presencia y niveles de elementos de interés nutricional, que pudieran afectar las cadenas alimentarias donde intervienen las especies estudiadas.
- Explorar la presencia patrones presentes en los datos químicos obtenidos, mediante técnicas quimiométricas.
- Proponer modelos quimiométricos que permitan certificar origen de futuras

muestras.

Hipótesis de Trabajo

Es posible establecer modelos quimiométricos de la procedencia geográfica de tres especies vegetales de importancia regional como forrajeras, teniendo en cuenta la composición mineral (elementos a nivel de vestigios), de material botánico analizado por técnicas analíticas.

2) Capítulo II

2.1 Introducción

La técnica de espectrometría de masas con plasma de acoplamiento inductivo (ICP-MS, o Inductively Coupled Plasma Mass Spectrometry), puede ser vista como una de las técnicas principales en el análisis elemental centrándose en la determinación de metales en concentraciones de ultratraza en una diversa variedad de muestras (Balcaen, L. et al. 2015). Entre las principales ventajas podemos mencionar que es una técnica multielemental, con alta sensibilidad para casi todos los analitos, con rango lineal de varios órdenes de magnitud, su posibilidad de acoplamiento a otras técnicas, de manera rápida. Es entonces una técnica óptima en la industria de alimentos, medioambiental, geológica, clínica, farmacéutica, química, petroquímica, nuclear o forense, permitiendo detectar y medir la mayoría de los elementos de la tabla periódica (Thomas, R. 2008).

Los elementos mostrados en color en la Fig. 2.1 pueden ser analizados por el ICP-MS con niveles de detección en o por debajo de los ppb (partes por billón). Los elementos que están en blanco no pueden ser medidos por ICP-MS ya que la ocurrencia de los isótopos es nula en la naturaleza.

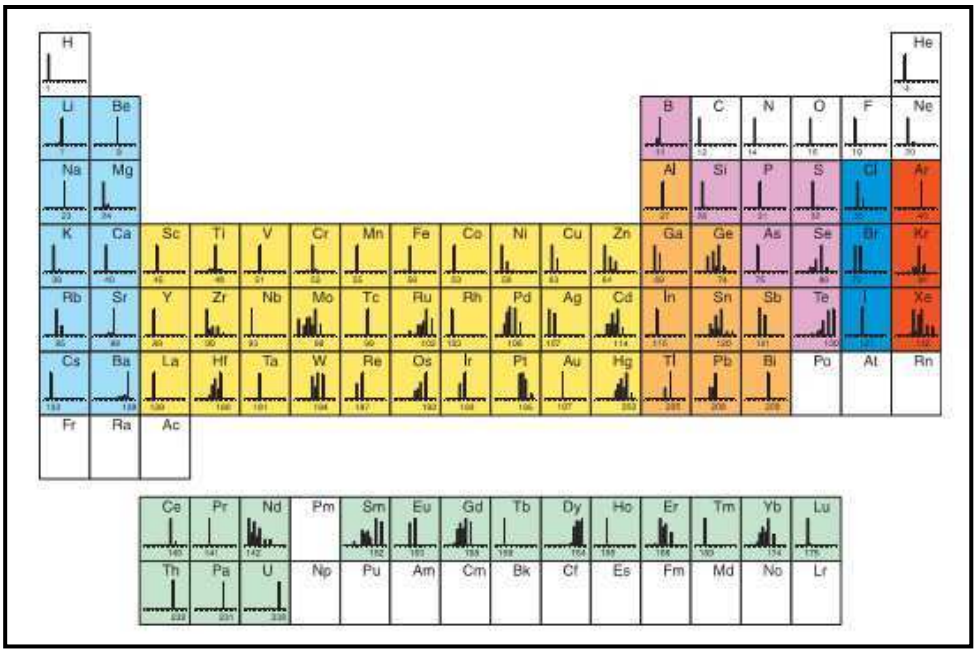


Fig. 2.1 Tabla periódica de elementos en la que se detallan en color los elementos que se pueden detectar mediante la técnica de ICP-MS

La aplicación del ICP-MS a la autenticación y trazabilidad de los alimentos ha aumentado rápidamente a lo largo de los años como consecuencia de tres factores principales: una mayor comprensión de la relación entre suelo y plantas desde el punto de vista de los minerales; una mayor comprensión de los procesos involucrados en las cadenas de producción de alimentos; y un fuerte aumento en la sensibilidad en los resultados analíticos de las instrumentaciones ICP-MS (Aceto, M. 2016, Gonzalvez, A. et al. 2009).

2.1 Principios del plasma de acoplamiento inductivo de argón

Un plasma es un gas ionizado que es macroscópicamente neutro, es decir, con el mismo número de partículas positivas y negativas. En contraste con una llama, es necesario suministrar una energía externa en forma de campo eléctrico para ionizar el gas y sostener el plasma, el cual, a su vez, transmitirá parte de esa energía a la muestra para atomizarla, ionizarla y excitarla. Los plasmas pueden clasificarse de

acuerdo con el tipo de campo eléctrico que se utiliza para crear y mantener el plasma. En el caso particular, del plasma acoplado inductivamente se obtiene mediante un campo de alta frecuencia a través de una bobina. Ese plasma por efecto Joule alcanza temperaturas alrededor de los 8000 °C (Hill, S. J. 2008).

La generación de plasma ocurre como consecuencia de un flujo de gas, Ar usualmente, sometido a la acción de un campo magnético oscilante, que es inducido por una corriente a alta frecuencia. Cuando se genera el plasma, los iones de Ar y los electrones libres que están presentes son acelerados siguiendo una trayectoria anular, ya que el campo magnético generado por la fuente de radiofrecuencias es alterno.

Por efecto Joule, en este caso de fricción iónica y electrónica, se consiguen energías altas, llegando a obtener temperaturas de hasta 8000 °C en las zonas de máxima intensidad de campo. En el seno del plasma se encuentran, coexistiendo: electrones (e^-), iones de argón (Ar^+), átomos de argón en estado fundamental (Ar^0), átomos de argón excitados (Ar*), moléculas de argón ionizadas (Ar_2^+), neutras (Ar_2^0) y excitadas (Ar_2*). Esto trae como consecuencia que se produzcan procesos de recombinación:

$$Ar + e^- \rightarrow Ar^+ + 2\,e^-$$

$$Ar^+ + e^- \rightarrow Ar^* + hv$$

$$Ar^* \rightarrow Ar^0 + hv$$

Debido a esto es que el plasma tiene aspecto de llama, aunque no lo sea, ya que no existen procesos de combustión química convencional. En la Fig. 2.2 se esquematizan los pasos desde la nebulización hasta el análisis de los iones.

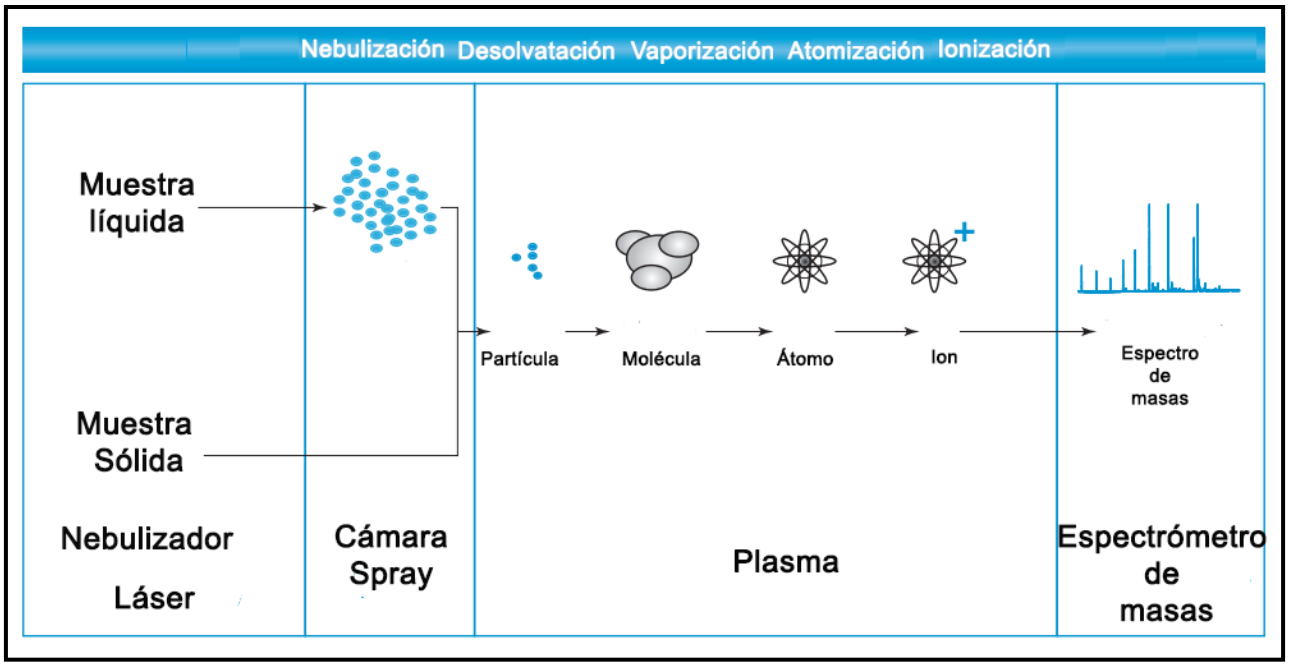


Fig. 2.2 Procesos químicos involucrados en un equipo ICP-MS

En la Tabla 2.1 se muestran los elementos que pueden ser excitados mediante el plasma de Ar.

Tabla 2.1. Elementos excitados con carga mono, divalente, trivalente y no detectable por ICP-MS.

Elementos excitados como M^+	Ar, Ag, Al, As, Au, B, Be, Bi, Br, C, Cd, Cl, Co, Cr, Cs, Cu, Dy, Er, Fe, Ga, Ge, H, Hg, I, In, Ir, K, Kr, Li, Mn, Mo, N, Na, Nd, Ni, O, Os, P, Pd, Pr, Pt, Pu, Rb, Re, Rh, Rn, Ru, S, S, Se, Si, Ta, Tb, Te, Th, Tl, Tm, U, W, Xe, Zn
Elementos excitados como M^+, M^{++}	Ba, Ca, Ce, Eu, Gd, Hf, La, Lu, Mg, Nb, Pb, Ra, Sc, Sm, Sn, Sr, Tc, Ti, V, Y, Yb, Zr
Elementos Indetectables por ICP-MS	F, He, Ne

2.2 Principios de la espectrometría de masas

Un espectrómetro de masas es un instrumento que produce iones y los separa de acuerdo con sus relaciones de masa/carga. La mayor parte de los iones que se estudian tienen una sola carga, de modo que la relación es simplemente el número de la masa del ion. La función del analizador de masas es similar a la de un

monocromador de un espectrómetro óptico. En el analizador de masas la dispersión depende de la relación masa-carga de los iones del analito y no de la longitud de onda de los fotones. Al igual que un espectrómetro óptico, un espectrómetro de masas contiene un transductor que convierte el haz de iones en una señal eléctrica que pueda ser procesada, almacenada en la memoria de una computadora y mostrada en una pantalla o almacenada en otros medios. A diferencia de la mayoría de los espectrómetros ópticos, los espectrómetros de masas requieren un complejo sistema de vacío para mantener una presión baja en todos los componentes, excepto en los sistemas para procesar la señal y la lectura. La presión baja asegura colisiones no frecuentes en el espectrómetro de masas para producir y conservar iones y electrones libres (Skoog, D. A. et al. 2008).

2.3 Componentes de un Equipo ICP-MS

2.3.1 Nebulizadores

La función del nebulizador es convertir a la solución en un aerosol que pueda ser transportado por una corriente gaseosa al plasma, introducir la muestra de tal manera que todos los procesos que ocurren en el plasma lo hagan de manera reproducible y finalmente producir gotas pequeñas. Junto con la cámara espray forman el sistema de inyección de muestra hacia la antorcha de plasma acoplado inductivamente (Dean, J. R. 2005, Thomas, R. 2008). Existen diversos tipos de nebulizadores, cada uno de ellos con sus ventajas y sus inconvenientes. En la Fig. 2.3 se muestra un esquema de un nebulizador (Dean, J. R. 2005).

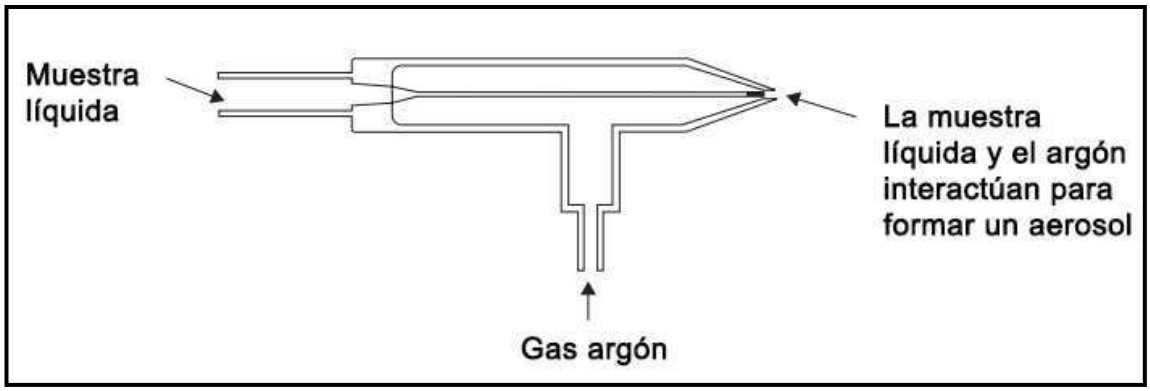


Fig. 2.3 Diagrama esquemático de un nebulizador neumático concéntrico

2.3.2 Cámara Spray

La función primaria de la cámara spray es eliminar grandes gotitas y permitir el paso de aquellas que pueden ser procesadas por el plasma para entrar en la interfase de acondicionamiento. Para lograr este objetivo, aparentemente tan simple, unos amplios rangos de procesos complejos ocurren que pueden ser beneficiosos o perjudiciales en el desempeño analítico (Hill, S. J. 2008). Podemos decir que los efectos de la cámara spray son: reducir la cantidad de aerosol que llega al plasma, disminuir la turbulencia asociada al proceso de nebulización y reducir el tamaño de las gotas de aerosol (Dean, J. R. 2005). El flujo de gas al entrar en la cámara spray sufre cambios en la dirección (180°), esto hace que las gotas grandes queden en las paredes y sean desechadas posteriormente. De esta forma, se asegura que las gotas pequeñas sean las que permanezcan y lleguen al plasma. En la cámara spray se pierde alrededor de 99% de la muestra en solución. En la Fig. 2.4 se presenta un esquema de una cámara spray (Dean, J. R. 2005).

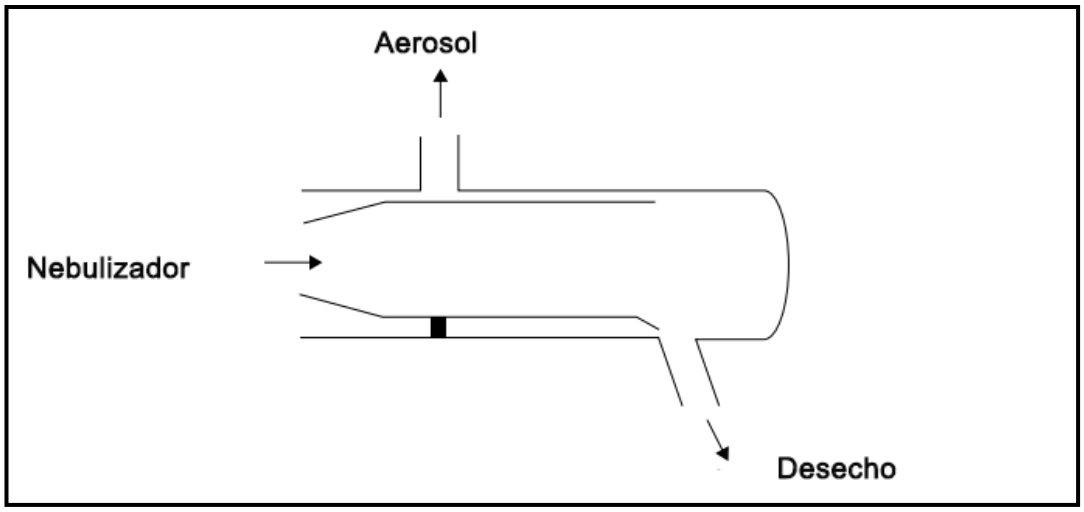


Fig. 2.4 Dibujo esquemático de una cámara de spray de doble paso

2.3.3 Antorcha

La antorcha de plasma consiste en tres tubos concéntricos, que comúnmente están hechos de cuarzo. En la Fig. 2.5 se observan el tubo exterior, el tubo intermedio y el inyector de muestra. La antorcha puede ser de una pieza, comúnmente conocida como el diseño de Fassel, en donde los tres tubos están conectados, o un diseño desmontable donde los tubos y el inyector de muestra están separados. El gas (usualmente argón) que es utilizada para formar el plasma (estado gaseoso) que pasa entre los tubos externos e intermedios a un caudal de aproximadamente 12-17 L/min. Un segundo flujo de gas (gas auxiliar) también a un caudal de 1 L/min trae la muestra, en la forma de finas gotas de aerosol, desde el sistema de inyección de muestra. El inyector de la muestra está a menudo hecho de otros materiales además del cuarzo, como alumina, platino y zafiro si necesitan utilizarse materiales corrosivos. En general, el argón es el más apropiado para usar para usar en los tres flujos, existen beneficios analíticos de usar mezclas de gases especialmente en el flujo del nebulizador. La antorcha de plasma está montada horizontalmente y situada en el centro de la bobina

de radiofrecuencia aproximadamente a 10-20 mm desde la interfase (Thomas, R. 2008).

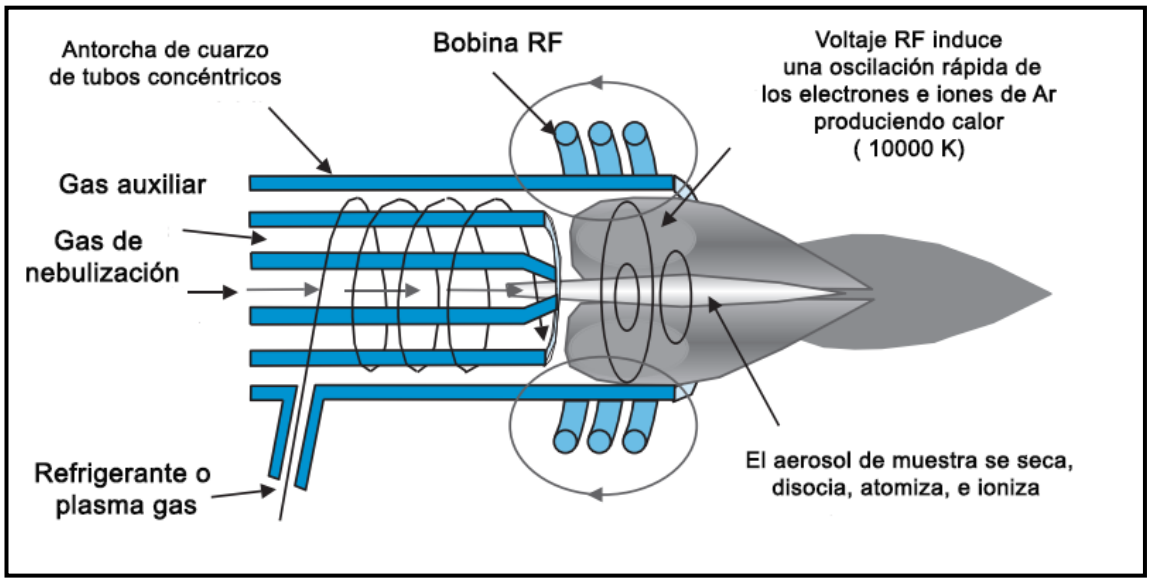


Fig. 2.5 Plasma generado por la antorcha en contacto con la interfase de vacío.

En la Fig. 2.6 podemos apreciar cómo se genera el plasma (Thomas, R. 2008):

- Un flujo tangencial de Ar pasa por la camisa externa de la antorcha.
- Se aplica potencia de radio frecuencia al load coil, produciendo un intenso campo electromagnético en su zona de influencia.
- Una chispa de alto voltaje aplicada sobre el argón produce electrones libres.
- Los electrones libres son acelerados por el campo de radio frecuencia produciendo colisiones e ionizando el Ar.
- El plasma de acoplamiento inductivo (ICP) se ha generado y se confina al final del lado abierto de la antorcha.

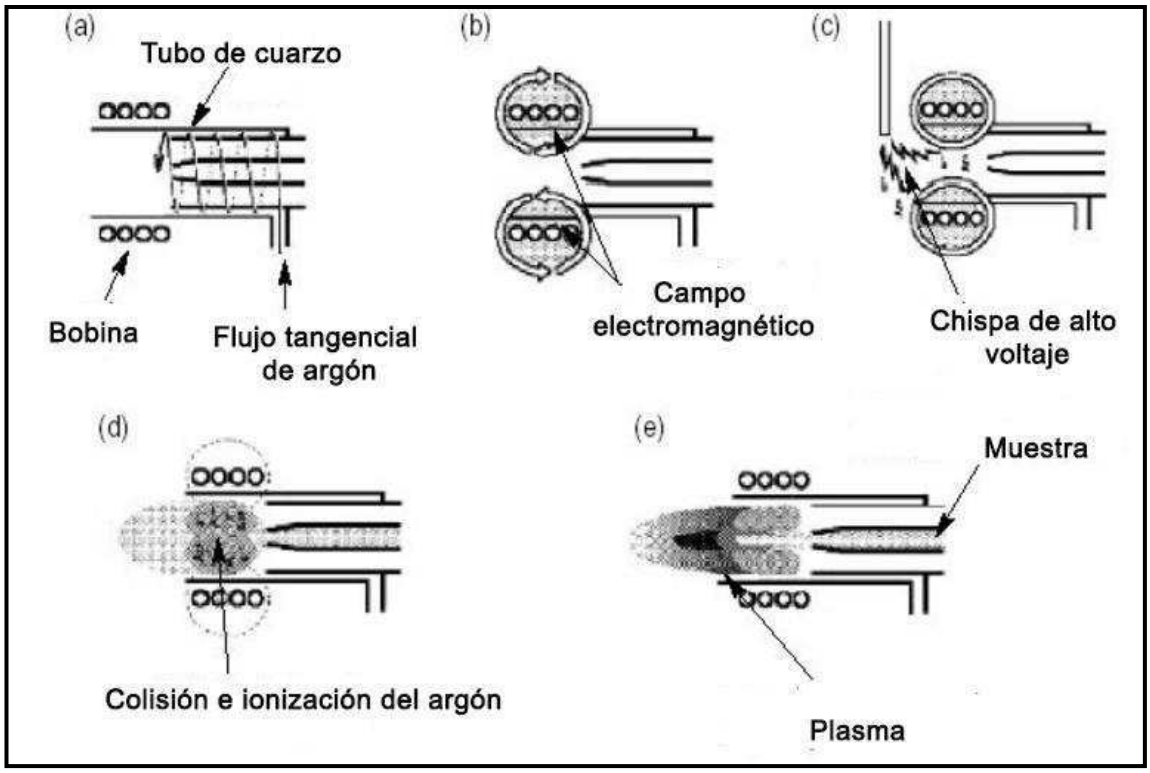


Fig. 2.6 Esquema de formación del plasma

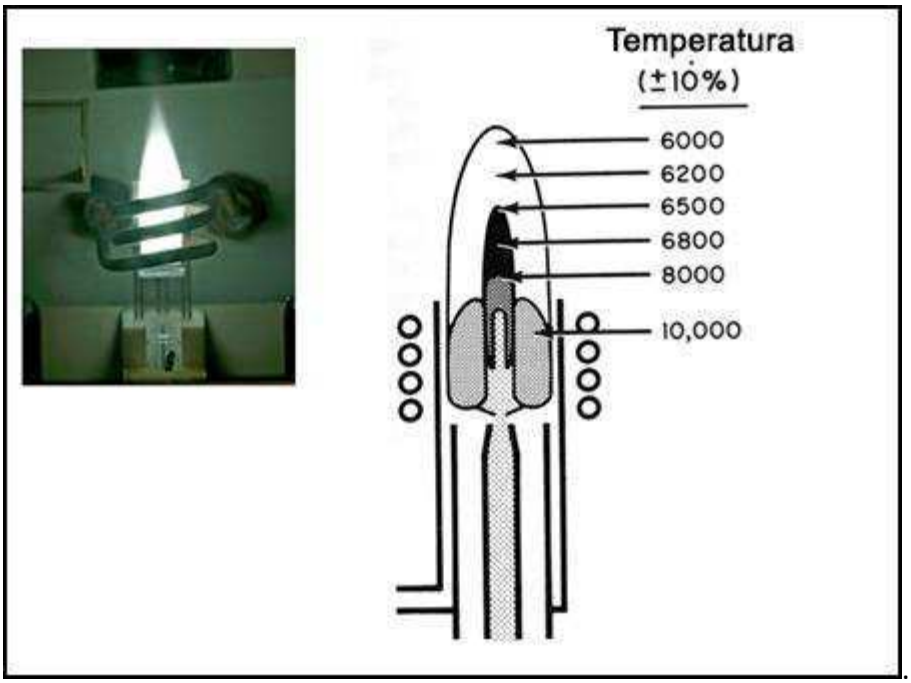


Fig. 2.7 Corte transversal de un plasma con el correspondiente perfil de temperaturas.

En la Fig. 2.7 se visualizan las temperaturas en varias zonas del plasma. En el momento en que los átomos de la muestra alcanzan el punto de observación, habrán permanecido unos 2 milisegundos a temperaturas comprendidas entre 4.000 y 10.000 K.

2.3.4 Interfase de Acondicionamiento

La interfase es necesaria ya que existe una diferencia de presiones entre el plasma y el sistema de detección. Mientras que el plasma opera a condiciones atmosféricas, el analizador de masas y el detector se encuentran a alto vacío. En la Fig. 2.8 se observa el lugar de la interfase en un equipo ICP-MS.

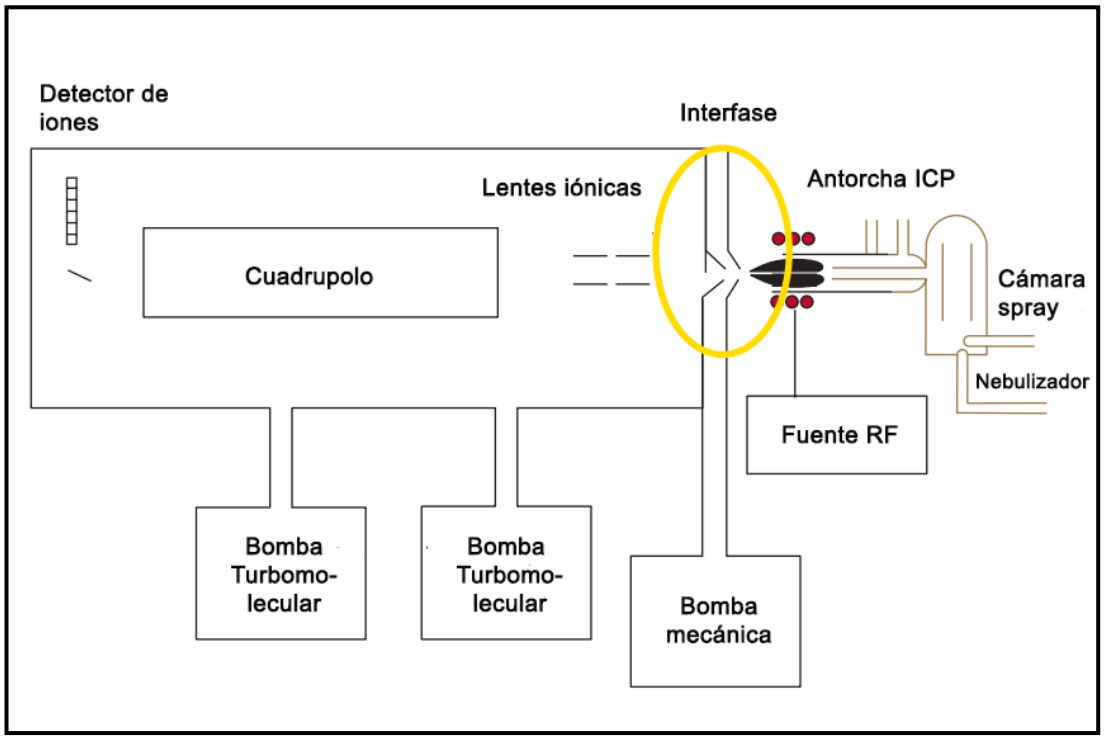


Fig. 2.8 Esquema de la interfase en el equipo ICP-MS

En esta etapa la idea es que el flujo de iones pase por una serie de cámaras con presiones cada vez menores. Consiste en dos conos, el de muestreo y el cono skimmer. La presión varía entre presión atmosférica (760 Torr) a una presión de 2 Torr que

existe en la zona entre los dos conos. Este vacío se genera por medio de una bomba rotatoria (Thomas, R. 2008). Se utilizan diversos materiales tales como Al, Cu, Ni y Pt, aunque el Ni es el mejor en la relación calidad/precio. El material sobre el que se construye debe tener una buena conductividad tanto térmica como eléctrica. Si se introducen materiales orgánicos, se debe utilizar conos de Pt ya que es menos susceptible de degradación que el Ni. En el caso de que se inyecte agua regia no se deben de utilizar los conos de Pt ya que esta es la única mezcla ácida que ataca al Pt.

El skimmer es más pequeño y menos robusto que el sampler, sobre todo en la punta, la cual es mucho más aguda y regular. Los conos skimmer se colocan generalmente entre 6 a 7 mm de distancia del de muestreo y también se fabrica con Ni o Pt. La punta de este cono tiene una repercusión directa sobre la sensibilidad del aparato.

2.3.5 Lentes Iónicas

La función de las lentes iónicas es transportar los iones procedentes del skimmer al espectrómetro de masas. En la Fig. 2.9 se presenta el lugar que ocupan las lentes iónicas en un equipo ICP-MS.

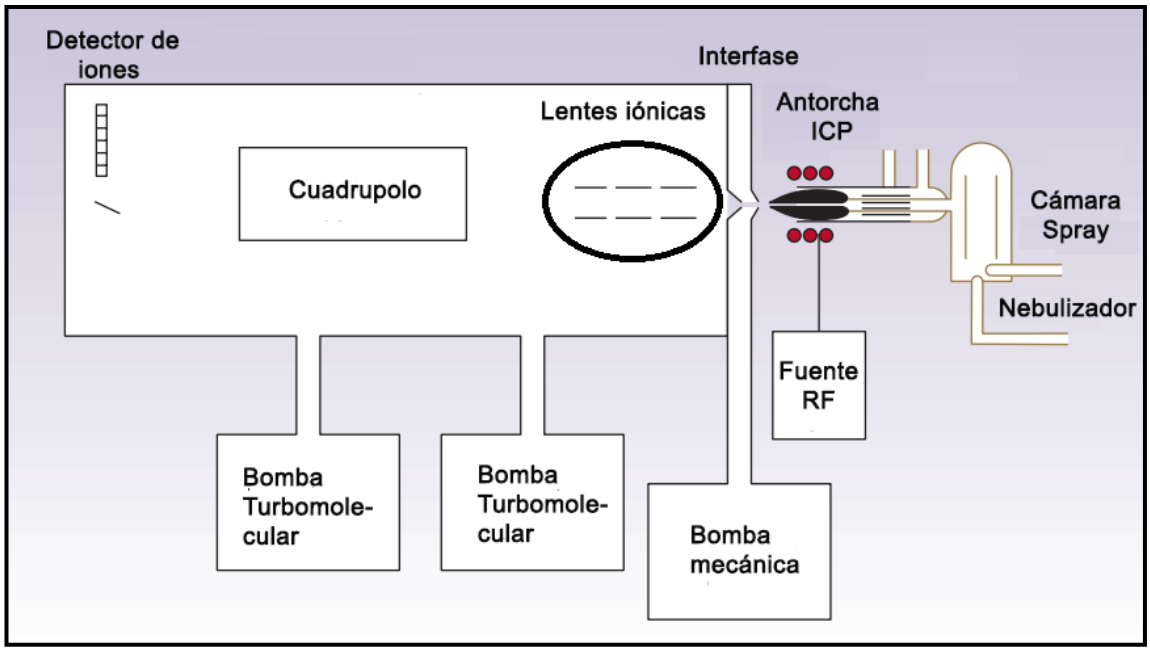


Fig. 2.9 Esquema de las lentes iónicas en el equipo ICP-MS

Una vez que los iones han entrado en la punta del skimmer es necesario extraer y enfocar los iones hacia el cuadrupolo. Las lentes de enfoque iónico son cruciales para la sensibilidad general del instrumento porque los iones dispersados no serán detectados. El enfoque de estos iones se logra sometiendo los iones cargados a campos eléctricos constantes. Estos campos eléctricos tienen un efecto sobre los iones y si la intensidad del campo es alta, los iones pueden acelerarse de tal forma que el tiempo en el cuadrupolo no es bueno para un análisis de masa eficaz (Hill, S. J. 2008).

2.3.6 Espectrómetro de masas cuadrupolar

La función del espectrómetro de masas cuadrupolar consiste en separar los iones en función de su relación carga/masa. En la Fig. 2.10 se presenta un esquema de un cuadrupolo (Vanhaecke, F. and Degryse, P. 2012).

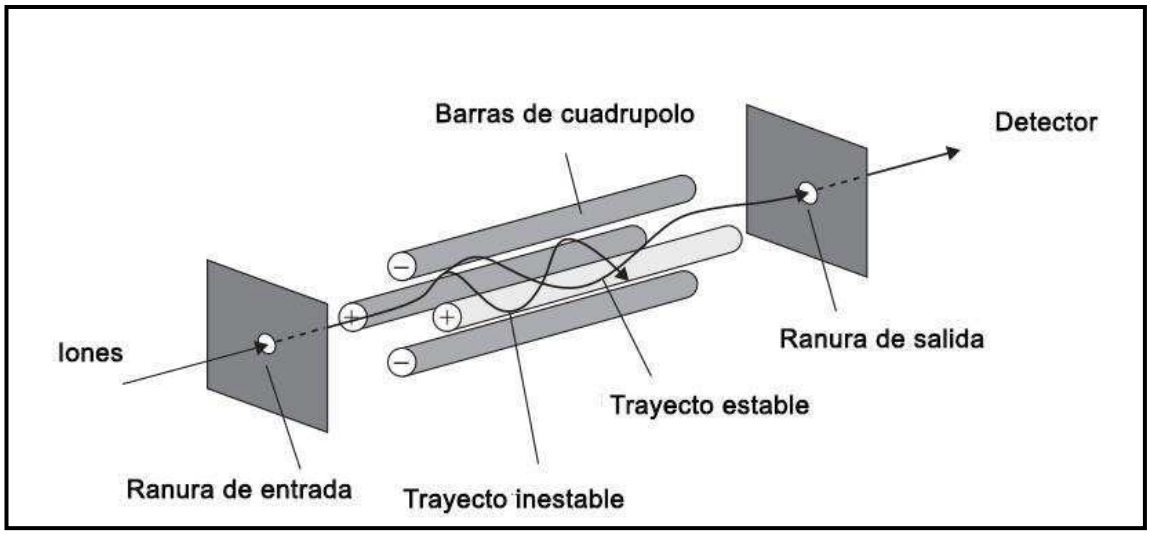


Fig. 2.10 Esquema del cuadrupolo en un equipo ICP-MS

Este tipo de analizador está compuesto de cuatro barras organizadas paralelamente como dos grupos de dos barras conectadas eléctricamente y equidistantes al eje. Una combinación de voltajes de radiofrecuencia y corriente directa son aplicados a cada par de barras, este arreglo simétrico permite producir campos hiperbólicos. Superficies diagonalmente opuestas están conectadas juntas a fuentes de voltajes de radiofrecuencia y de corriente directa. Los iones se extraen de la fuente de iones y son acelerados (5-15 V) dentro del espacio central formado por el cuadrupolo a lo largo del eje longitudinal hacia el detector.

Las trayectorias de los iones a través del espacio central de las barras son complicadas, y para cada par de voltajes de corriente directa y de corriente alterna solo iones de un valor m/z específico evitaran colisionar con las barras y pasaran el filtro cuadrupolo a través del eje Z hasta alcanzar el detector, todos los demás iones chocarán con las superficies del cuadrupolo a estos valores de radiofrecuencia y de corriente directa. Tales trayectorias se les denominan inestables ya que los iones tienen caminos inestables. Los iones con caminos estables son aquellos que se

encuentran dentro de las barras. En la Fig. 2.11 se presenta el esquema de funcionamiento del cuadrupolo (Vanhaecke, F. and Degryse, P. 2012).

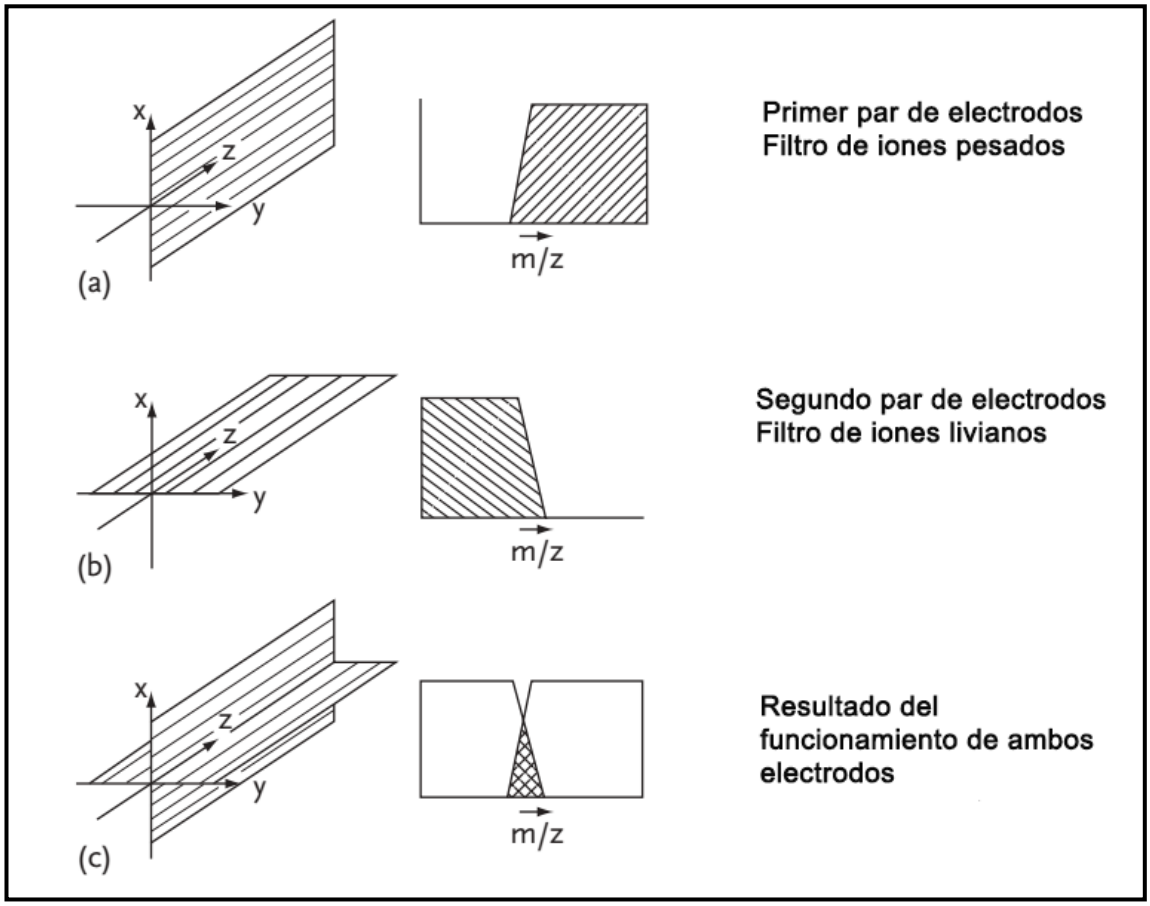


Fig. 2.11 Esquema de la acción del filtro de masas cuadrupolar

2.3.7 Detectores

El más utilizado es un electrón multiplicador que tiene un efecto muy parecido al fotomultiplicador. Es un tubo de vidrio abierto con un cono en una terminación. Para la detección de iones positivos, el cono es sometido a un alto potencial negativo (aproximadamente -3kV). En la Fig. 2.12 se presenta un esquema del funcionamiento de un tubo fotomultiplicador (Dean, J. R. 2005).

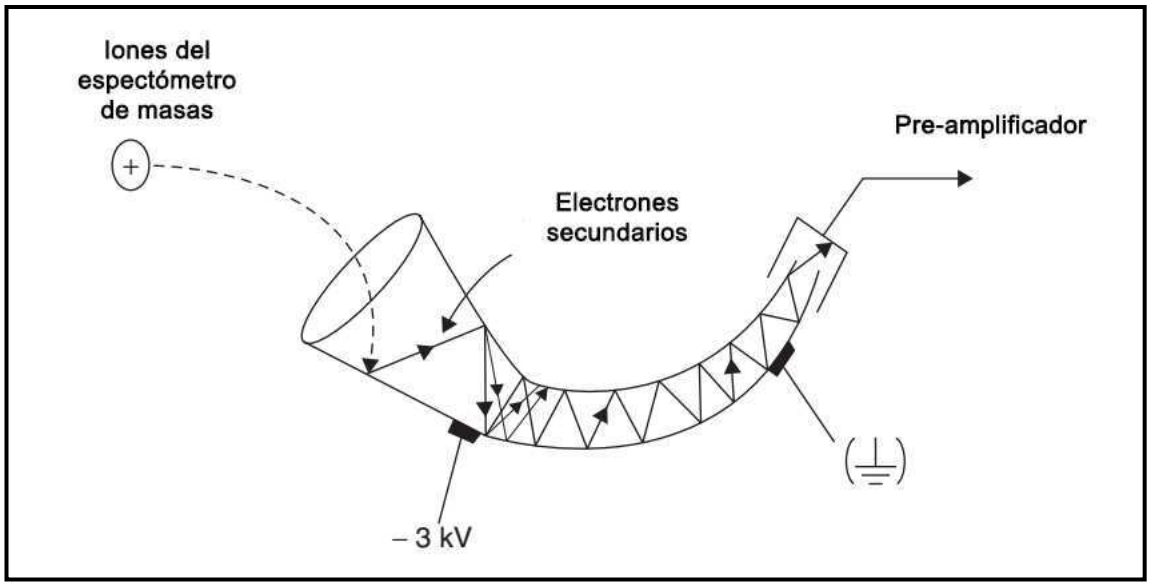


Fig. 2.12 Esquema de funcionamiento de un tubo fotomultiplicador

Cuando los iones salen del analizador de masas, son atraídos por el potencial negativo del cono. Cuando los iones chocan con su superficie, se originan uno o más electrones secundarios. El potencial dentro del tubo varía continuamente con la posición, tal que los electrones secundarios se mueven hasta llegar a otra zona donde se originan otros electrones secundarios, y así sucesivamente. El tubo está cubierto por un material semiconductor. La vida media del multiplicador está determinada por la carga total acumulada.

2.3.8 Sistemas de vacío

El plasma se encuentra a presión atmosférica mientras que el espectrómetro de masas requiere el movimiento de iones sin colisión, o bajo un sistema de vacío. La estrategia de vacío para el ICP-MS es reducir gradualmente la presión en estadios separados por una técnica que se le denomina bombeo diferencial. El sistema tiene tres orificios (sampler, skimmer y orificio de bombeo diferencial). En el primer estadio se utiliza un bombeo mecánico. La velocidad de bombeo es baja.

En las lentes iónicas, la presión es de 10^{-5} Torr. Virtualmente todo el gas que viene desde el plasma a través del skimmer es evacuado. El tercer estadio es de 10^{-7} Torr o menos para la zona del cuadrupolo y detector. Las bombas rotatorias o mecánicas se utilizan para la interfase. Se fundamenta en el giro de un rotor dentro de un cilindro de metal o estator. Otro tipo de bombas son las bombas turbomoleculares, las cuales originan alto vacío en poco tiempo. Se trata de una turbina que gira a altas velocidades (30.000 rpm) dentro de un cilindro hueco. El gas de la cámara fluye a la bomba y es comprimido por la turbina. Aquí se utiliza líquido refrigerante.

2.4 Interferencias de la técnica ICP-MS

En el trabajo con ICP-MS existen dos grupos de interferencias principales: espectroscópicas y no espectroscópicas (Hill, S. J. 2008). Las interferencias espectroscópicas pueden ser:

- Solapamientos isobáricos
- Iones Poliatómicos
- Iones con carga doble

Las interferencias no espectroscópicas se subdividen en efectos de supresión y efectos físicos causados por altos sólidos disueltos totales. La extensión de los problemas de interferencia está relacionada con la naturaleza de la matriz de la muestra, por lo tanto, pueden ser minimizados con una cuidadosa preparación de ésta. En este trabajo se detalla la explicación de las interferencias espectroscópicas debido a que es el más frecuente en el tipo de muestras analizadas.

2.4.1 Solapamientos isobáricos

Dos elementos que tienen isótopos de igual masa producen lo que se conoce como solapamiento isobárico. En general, las dos masas pueden diferir por una cantidad ínfima, pero la resolución 0.005 m/z no puede ser resuelta por un cuadrupolo del equipo de ICP-MS. Para discriminar las masas hay que utilizar un sistema de resolución con un doble enfoque de masas. Como regla general, las masas impares tienen solapamientos, no así las masas pares. Hoy en día, varios softwares de ICP incorporan estas correcciones. Pero, aun así, realizando estas correcciones existe un grado de error.

2.4.2 Iones Poliatómicos

Desde un punto de vista práctico, este tipo de interferencias son más graves que los solapamientos isobáricos. Los iones resultan de una combinación de corta vida de dos o más especies atómicas por ejemplo ArO^+, Ar, H y O son las especies que predominan en el plasma y pueden unirse entre sí o con otros elementos de la matriz, con elementos que están formando parte de solventes o ácidos que se utilizan en el acondicionamiento de la muestra (N, S, y Cl). Cuando las masas son superiores a 82, es posible la formación de iones poliatómicos. En esto influye el sistema nebulizador, los parámetros del plasma y la geometría de extracción, aunque los factores más importantes son tanto la matriz como los ácidos utilizados durante el análisis (Dean, J. R. 2005).

2.4.3 Iones con carga doble

En el análisis multielemental con ICP la mayoría de los iones son de carga única, aunque algunas especies son de carga múltiple. Esto está íntimamente vinculado a la segunda energía de ionización del elemento y las condiciones de equilibrio del plasma. Si estos elementos tienen una energía de ionización más baja que la primera del Ar, es muy posible la formación de carga 2+. Corresponden a este grupo, los alcalinotérreos, metales de transición y tierras raras. En general, la presencia de iones con carga doble implica una perdida en la sensibilidad para la especie con carga simple, ya que se produce un solapamiento isotópico entre ambas (Dean, J. R. 2005).

2.5 Referencias Bibliográficas

Aceto M. 2016. 8 - The Use of ICP-MS in Food Traceability A2 - Espiñeira, Montserrat. In: Advances in Food Traceability Techniques and Technologies. Woodhead Publishing. p. 137-164.

Balcaen L, Bolea-Fernandez E, Resano M, Vanhaecke F. (2015) Inductively coupled plasma – Tandem mass spectrometry (ICP-MS/MS): A powerful and universal tool for the interference-free determination of (ultra)trace elements – A tutorial review. Analytica Chimica Acta.894:7-19.

Dean JR. (2005) Practical Inductively Coupled Plasma Spectroscopy: John Wiley & Sons.

Gonzalvez A, Armenta S, de la Guardia M. (2009) Trace-element composition and stable-isotope ratio for discrimination of foods with Protected Designation of Origin. TrAC Trends in Analytical Chemistry.28:1295-1311.

Hill SJ. (2008) Inductively Coupled Plasma Spectrometry and its Applications: John Wiley & Sons.

Skoog DA, Crouch SR, Holler FJ. (2008) Principios de analisis instrumental / Principles of Instrumental Analysis: Cengage Learning Latin America.

Thomas R. (2008) Practical Guide to ICP-MS: A Tutorial for Beginners, Second Edition: CRC Press.

Vanhaecke F, Degryse P. (2012) Isotopic Analysis: Fundamentals and Applications Using ICP-MS: John Wiley & Sons.

3) Capítulo III

3.1 Introducción

La Quimiometría puede ser brevemente descripta como el uso de ciertas herramientas matemáticas, estadísticas y de aprendizaje automático en el proceso de medida químico. Es una consecuencia en el cambio de los datos obtenidos con la emergencia de nuevas técnicas de análisis químico, como así también de los microprocesos. Su progreso en el siglo XXI se ha debido principalmente a la mejora de los programas estadísticos utilizados para ello (Kumar, N. et al. 2014).

Dentro de la quimiometría el reconocimiento de patrones ha ocupado un lugar importante. Algunas definiciones al respecto son (Brereton, R. G. 2015):

"El reconocimiento de patrones estadístico es un término utilizado para cubrir todas las etapas de una investigación desde la formulación del problema, la recolección de los datos, la discriminación y la clasificación, hasta la evaluación de los resultados y la interpretación." Webb

"El principal objetivo del reconocimiento de patrones es la clasificación supervisada y no supervisada". Jain

Los primeros pioneros en el reconocimiento de patrones han ampliado la definición a casi cualquier enfoque computacional utilizado para determinar patrones o relaciones entre objetos, pero nosotros nos centraremos en aquellas que se relacionan a una especie de clasificación. Una pregunta que nos podría surgir es si el título "reconocimiento de patrones" debería ser cambiado por el de "clasificación". Brereton (Brereton, R. G. 2015)

A su vez podemos definir el aprendizaje automático como una rama de la inteligencia artificial y su principal objetivo es desarrollar técnicas que permitan a las computadoras aprender. ¿Qué es lo que aprende un algoritmo? El aprendizaje se realiza sobre una base de datos, por lo que podríamos decir que se aprenden patrones subyacentes en un set de datos.

El aprendizaje automático se encuentra en la intersección entre la programación, la ingeniería y la estadística. Puede ser aplicada en variados y diversos campos desde la política a las ciencias de la tierra. Es una herramienta que puede ser aplicada a diversos problemas. Todo campo que pueda interpretar y obtener datos puede beneficiarse de las técnicas de aprendizaje automático (Harrington, P. 2011).

3.2 Datos multivariados

Los datos multivariados se presentan en una matriz de datos que consiste en *n* filas, *m* columnas, y cada celda conteniendo un valor numérico. Cada fila corresponde a un objeto, en primera instancia, una muestra; y cada columna representa una característica de la muestra (una variable, una medida que es cuantificable). Llamamos a esta matriz, *X*, con el elemento x_{ij} en la fila *i* y la comuna *j*. El vector columna contiene los valores de la variable *j* para todos los objetos; y el vector fila x_i^T es el vector traspuesto y contiene todas las variables para el objeto *i* (Fig 3.1).

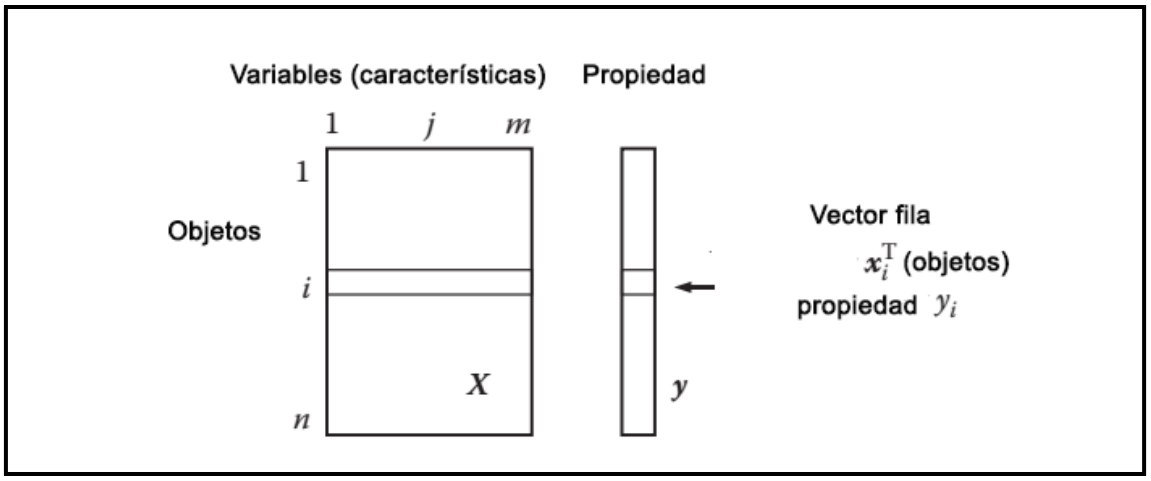


Fig. 3.1 Matriz de variables X y un vector de propiedades y. La propiedad puede ser continua (una propiedad física, química, biológica o tecnológica), como también, valores discretos o una variable categórica.

Geométricamente, cada objeto puede ser considerado como un punto en espacio m-dimensional con las coordenadas dadas por las m variables. Muchos métodos en el análisis de datos multivariado pueden ser entendidos fácilmente mediante una representación geométrica de una matriz de dos dimensiones. Un concepto importante es considerar la distancia entre objetos como una medida de la similitud de los objetos. Bajo el concepto de distancia, los agrupamientos de objetos y datos extremos presentes pueden ser detectados (Varmuza, K. and Filzmoser, P. 2009).

3.3 Manipulación de datos

Muchos estimadores estadísticos descansan en la idea de la simetría de la distribución de los datos. Por ejemplo, la desviación estándar puede ser severamente incrementada si la distribución de datos está sesgada. Entonces, a menudo se recomienda realizar una primera transformación de los datos para una mejor simetría. Desafortunadamente, esto debe hacerse para cada variable por separado, porque no es seguro que la misma transformación matemática será útil para las demás variables (Varmuza, K. and Filzmoser, P. 2009).

3.3.1 Centrado y Escalado

El centrado y el escalado usualmente se aplican después de una transformación de datos. Todas las columnas de la matriz X tienen media cero (centrado) y la misma variancia (escalado). Dependiendo del análisis, sólo el centrado, pero no el escalado se aplica. Después del centrado de la media, la variable tiene una media de cero; y las distancias entre los puntos (muestras) se mantienen sin cambio.

$$x_{ij}\ (centrado\ de\ la\ media) = x_{ij}\ (original) - x_j$$

El escalado de variancia estandariza cada variable *j* por su desviación estándar sj, usualmente, esto se combina con una centrado de media y se llama autoescalado (o transformación z)

$$X_{ij}(autoescalado) = \frac{x_{ij}\ (original) - x_j}{s_j}$$

Los datos autoescalados tienen una media de cero y una variancia de uno, de este modo, todas las variables tienen un peso estadístico igual. El autoescalado es el método más usado para pre-procesar los datos en Quimiometría. (Varmuza, K. and Filzmoser, P. 2009).

3.3.2 Normalización

La normalización se refiere a la transformación de filas de la matriz, *X*. La transformación usual puede ser la normalización a una suma constante de variables (para datos de concentración), una constante máxima de valor de variables (para datos de masas espectrales) o normalización a un vector de longitud constante (Varmuza, K. and Filzmoser, P. 2009).

3.4 Análisis Exploratorio de Datos

3.4.1 Análisis de Componentes Principales (PCA)

El análisis de componentes principales (PCA) puede ser considerado como "la madre de todos los métodos en el análisis de datos multivariados" (Varmuza, K. and Filzmoser, P. 2009). El objetivo de PCA es la reducción de dimensiones y es el método más frecuentemente aplicado para el cálculo de variables latentes lineales (componentes). PCA puede ser visto como un método para calcular un nuevo sistema de coordenadas formado por variables latentes, que son ortogonales. Estas primeras variables son las que más información aportan de todo el set de variables. Las variables latentes de PCA representan de manera óptima las distancias entre los objetos altas dimensiones en el espacio variable. La distancia de los objetos que se considera como una similitud inversa de los objetos. PCA considera todas las variables y acomoda la estructura total de los datos; se trata de un método exploratorio de análisis de datos (aprendizaje no supervisado) y se puede aplicar a cualquier matriz X.

La reducción de dimensiones por PCA es usada principalmente para:

- Visualización de los datos multivariados por gráficos de dispersión.
- Separación de las variables relevantes (descripto por unas pocas variables latentes) del ruido.
- Combinación de varias variables que caracterizan un proceso químico-tecnológico en unas pocas variables.

El PCA es exitoso para un conjunto de datos con variables correlacionadas como es usual en datos químicos. Las variables constantes o variables altamente correlacionadas no causan problemas para la PCA; a pesar, de que los datos extremos influyen en los resultados, entonces el escalado se vuelve importante. En la Fig. 3.2 se

presenta el esquema de análisis matricial realizado en un análisis de componentes principales (Varmuza, K. and Filzmoser, P. 2009).

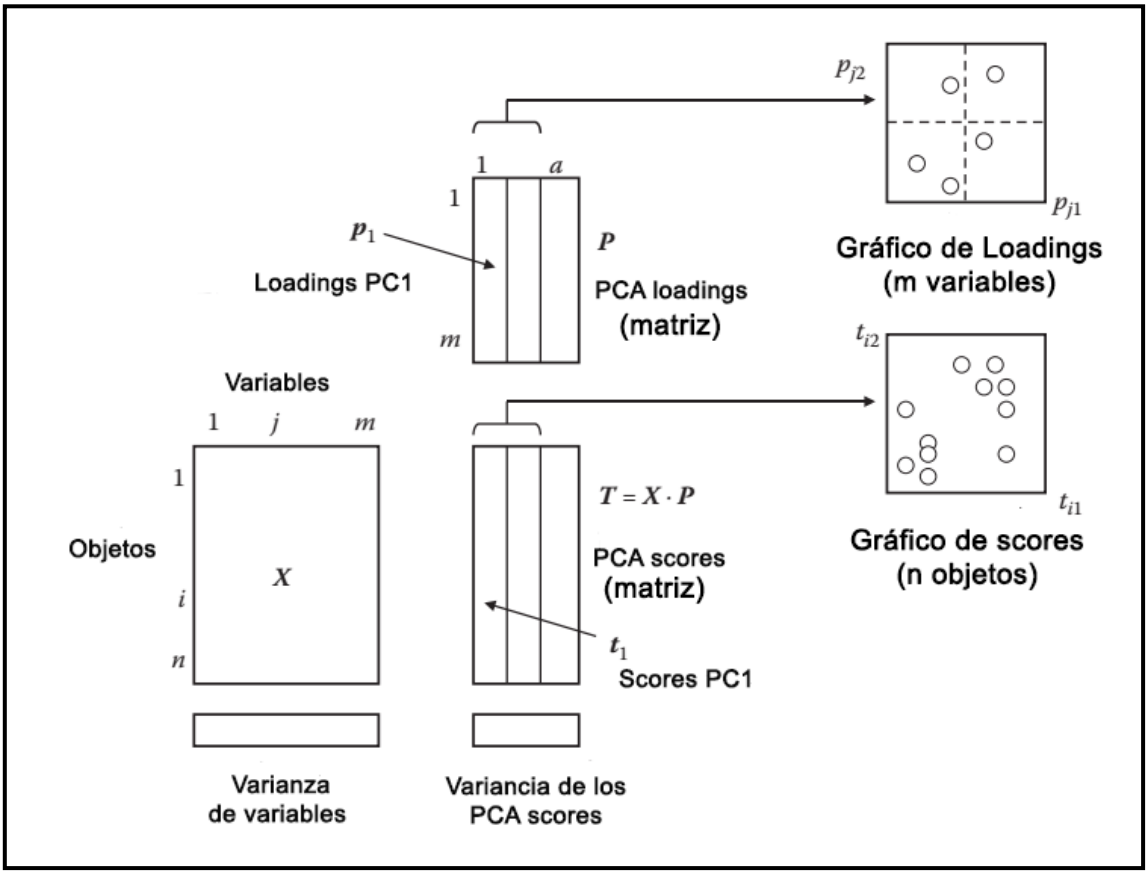


Fig. 3.2 Esquema de análisis matricial realizado por PCA.

La dirección en un espacio variable que mejor conserva las distancias relativas entre los objetos es una variable latente que tiene varianza máxima de los scores (éstos son los valores de los datos proyectados sobre la variable latente). Esta dirección es llamada por definición, el primer componente principal (PC1). Se define como vector loading

$$p_1 = (p_1, p_2, \ldots, p_m)$$

siendo *m* el número de variables.

En quimiometría, la letra p es ampliamente utilizada para loadings en PCA. Es común también normalizar la longitud de los vectores a 1; esto significa que $p_1^T p_1 = 1$.

Los correspondientes scores (proyección coordinada de los objetos, se denota con *t*) son combinaciones lineales de los loadings y las variables. Formalmente, el objeto i, definido por el vector xi con elementos de i a m, los scores t_{i1} para la PC1 es:

$$t_{i1} = x_{i1}p_1 + x_{i2}p_2 + \ldots + x_{im}p_m = x_i^T . p_1$$

La segunda parte de la ecuación expresa la proyección ortogonal de los datos en las variables latentes. Para los n objetos, dispuestos como filas en la matriz X, el vector de scores, t1, de la PC1 es obtenida por

$$t_1 = X . p_1$$

Todos los vectores ordenados en columnas en la matriz P, y todos los vectores scores en la matriz de scores, T

$$\mathbf{T} = \mathbf{X}.\mathbf{P}$$

Los scores tienen una propiedad matemática importante. Ellos son ortogonales entre sí, y como los scores están usualmente centrados, cualesquiera dos vectores no están correlacionados, resultando en un coeficiente de correlación cero. Ninguna otra rotación del sistema de coordenadas excepto el análisis de componentes principales tiene esta característica.

$$t_j^T . t_j = 0 \qquad i, j = 1, \ldots, m$$

La matriz *X* puede ser reconstruida desde los scores de PCA, *T*. Generalmente, unas pocas PCs son utilizadas correspondientes a la estructura principal de la base de datos. Esto resulta en una matriz aproximada con ruido reducido. Si todas las posibles PCs fuesen utilizadas, el error (residual) matricial E sería cero.

3.4.1.1 Número de componentes principales

El principal objetivo de la PCA es la reducción de dimensiones; que significa explicar la mayor variabilidad posible con la menor cantidad posible de componentes.

Si las correlaciones entre variables son pequeñas, no hay reducción de dimensiones posible sin la pérdida severa de variancia (información potencial).

La variancia de los scores de PCA –preferentemente dados en porcentaje total de variancia – son indicadores importantes. En un gráfico de scores, usando las dos primeras componentes principales, más del 70% de la variancia se preserva, el gráfico da una buena imagen de la matriz de datos multidimensional. Si más del 90% de la variancia total se mantiene, la representación dimensional es excelente, y la mayoría de las distancias entre scores reflejan bien la distancia en el espacio multidimensional. Si la suma de variancias de la PC1 y la PC2 es pequeña, gráficos adicionales, por ejemplo, de PC2 vs PC3, y de PC3 vs PC4 proveen información adicional y ayudan a reconocer la estructura de los datos.

Para estimar el número óptimo de componentes principales se puede recurrir a varias técnicas. El más usual es la variancia de los scores de la PCA vs el número de PCs. De acuerdo con la definición, la PC1 debe tener la variancia más grande, y las variancias irán decreciendo a medida que aumenta el número de componentes principales (Varmuza, K. and Filzmoser, P. 2009). El gráfico resultante se denomina gráfico de sedimentación y se detalla en la Fig. 3.3 (Varmuza, K. and Filzmoser, P. 2009).

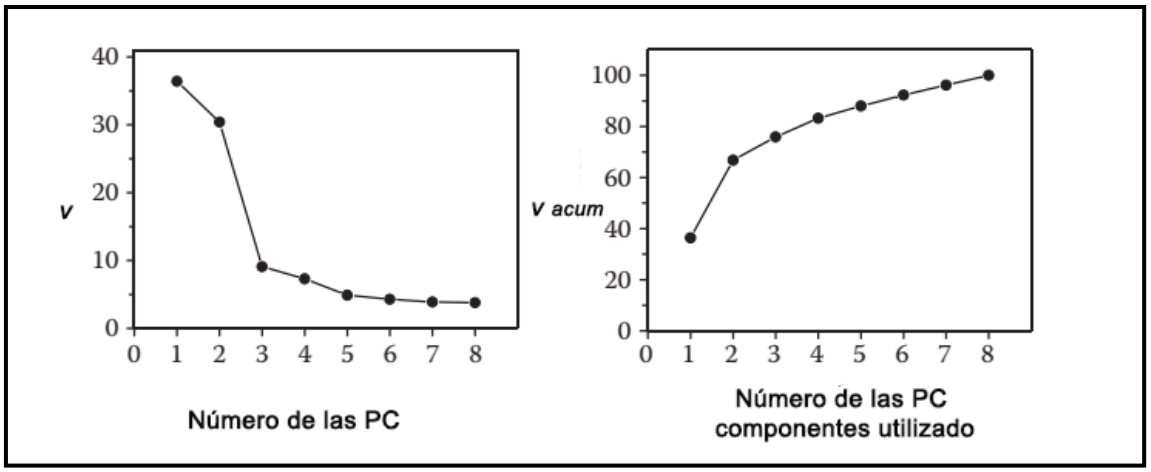


Fig. 3.3 Gráfico de sedimentación para un conjunto de datos artificial con 8 variables, siendo *v* la variancia de los scores PCA y v_{acum} la variancia acumulada de los scores de PCA

La variancia acumulada, v_{acum}, de los scores muestra cuanto de la variancia se conserva por las componentes de la PCA. Para las variables autoescaladas, cada variable tiene una variancia de 1, y una variancia total de m, el número de variables. Para estos datos, una regla de oro es utilizar las componentes de variancia mayor a 1. El número de componentes de PCA con variancia más grande que 0 es igual al rango de la matriz de covariancia de los datos (Varmuza, K. and Filzmoser, P. 2009).

3.4.2 Centrado y Escalado en el Análisis de Componentes Principales

Los resultados del análisis de PCA son fuertemente dependientes de la matriz de origen. La Fig. 3.4 muestra este efecto en un gráfico de dos dimensiones (Varmuza, K. and Filzmoser, P. 2009).

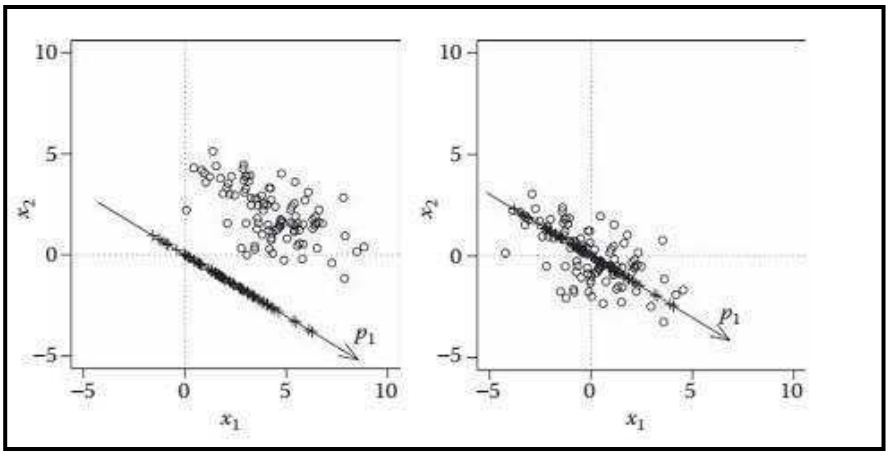


Fig. 3.4 Efecto del centrado de medias en el análisis PCA. En el gráfico de la izquierda los datos no están centrados al origen; en consecuencia, los scores no están centrados. El gráfico de la derecha muestra los datos centrados al origen.

En la imagen de la izquierda de la Fig 3.4 se observa el gráfico de scores de un análisis de componentes principales en el que no se ha realizado un centrado de los datos. La dirección de la PC1 se observa a lo largo de la nube de puntos, y nos indica la máxima variabilidad. Los scores de la PC1 son proyecciones ortogonales de los datos en esta dirección, y tienen una proyección positiva. Los scores no están centrados y esto no tiene consecuencia directa en la variabilidad de los datos, pero tendrá consecuencias posteriores cuando se realice una reducción dimensional con menor número de componentes principales. En el gráfico de la derecha, los scores se encuentran centrados (Varmuza, K. and Filzmoser, P. 2009).

Otro aspecto importante de la preparación de los datos para PCA es el autoescalado de los datos. Los resultados de PCA pueden cambiar si los datos originales (centrados con respecto a la media), o si los datos fueron previamente autoescalados (Varmuza, K. and Filzmoser, P. 2009). En la Fig 3.5 se muestran los efectos del escalado en un análisis de PCA (Varmuza, K. and Filzmoser, P. 2009).

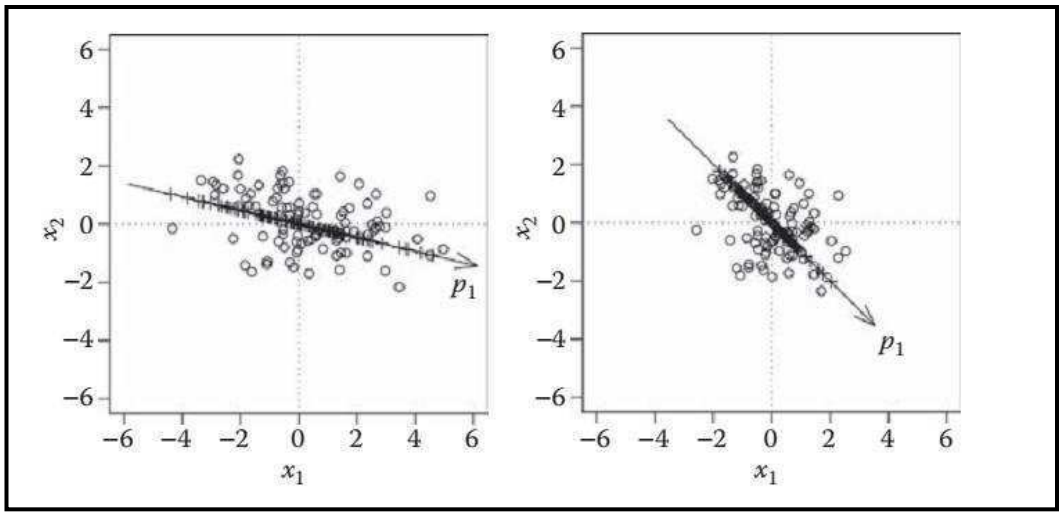


Fig. 3.5 Efecto del autoescalado en el análisis PCA. En el gráfico de la izquierda los datos no están escalados, sólo centrados. En el gráfico de la derecha están centrados y escalados.

En la primera imagen (Fig 3.5) se observan los resultados de una PCA con variables centradas con respecto a la media (las variancias de x_1 y x_2 son diferentes). En la segunda imagen los datos fueron autoescalados de tal forma que x_1 y x_2 ahora tienen la misma variancia (Varmuza, K. and Filzmoser, P. 2009).

A veces el escalado de los datos tiene un efecto indeseable, porque cada variable tendrá el mismo peso o importancia para la PCA. Entonces, las variables que esencialmente incluyen ruido serán tan importantes como las variables que reflejan la verdadera variabilidad. El análisis de PCA no puede distinguir entre la información importante y la no importante, y tratará de expresar tanta variabilidad como sea posible –también la variabilidad causada por el ruido. En estos casos, los datos no deberían ser escalados de manera de mantener la importancia original de las variables (Varmuza, K. and Filzmoser, P. 2009).

3.4.3 Métodos de Agrupamiento

Los métodos de agrupamiento son herramientas no supervisadas dentro del aprendizaje automático, que divide los datos en conglomerados o grupos de objetos

similares. Son más utilizados en la exploración de datos más que en la predicción, además de proveer una visión de los agrupamientos dentro de los datos (Lantz, B. 2015).

Usualmente se refiere a los objetos (en el espacio variable), pero también se utiliza para variables (en el espacio de las variables), o también para ambos, variables y objetos simultáneamente. Hablando en términos de objetos, el análisis de agrupamiento intenta identificar grupos concentrados (conglomerados) de objetos, mientras ninguna información de los objetos está disponible, y usualmente ni siquiera el número de conglomerados es conocido.

La tarea de encontrar grupos similares de objetos presume que esa estructura de grupos está inherente en el conjunto de datos. En general, no se asume que un objeto pertenece a un único grupo, pero puede ser parte de dos o más grupos. Entonces, los métodos de agrupamiento que realizan una separación de objetos en grupos separados no siempre dan la solución deseada. Por esta razón, muchos algoritmos de conglomerados que se proponen en la literatura no sólo operan de manera distinta, sino que incluso actúan sobre principios diferentes. Los métodos más importantes son:

- Métodos de partición: cada objeto es asignado a un grupo.
- Métodos jerárquicos: objetos y particiones están dispuestos en una jerarquía. Una representación gráfica es el dendrograma. Permite determinar manualmente el número óptimo de conglomerados como de relaciones jerárquicas entre los diferentes grupos de objetos.

- Métodos basados en modelos: los diferentes conglomerados se suponen que siguen un determinado modelo, como una distribución normal multivariante con una cierta media y covariancia.

El resultado del procedimiento de cualquier análisis de conglomerados es la asignación de los objetos a los grupos, donde los objetos dentro de los grupos se suponen que son similares entre ellos, y los objetos de diferentes grupos se suponen que son diferentes. Esto trae aparejado otras cuestiones:

- La cercanía entre objetos debe ser medida, lo que se realiza mediante una distancia o una medida de similitud.
- Como el número correcto de conglomerados es desconocido, la validación de la medida de conglomerados necesita ser consultada para la evaluación del análisis.

Usualmente uno no puede pretender una única solución del análisis de conglomerados. El resultado depende de la medida de distancia utilizada, el algoritmo de agrupamiento, y los parámetros elegidos, generalmente desde las condiciones iniciales. El éxito de la técnica de conglomerados está determinado si los conglomerados encontrados pueden ser asignados a un grupo que sean importantes para el análisis, o no. La aplicación de métodos no supervisados es a menudo recomendable como un paso inicial en la evaluación de los datos para tener una visión de la estructura de los datos (para detectar agrupamientos o valores atípicos) que pueden ser importantes para la posterior aplicación de un método clasificatorio (Varmuza, K. and Filzmoser, P. 2009).

En el análisis de conglomerados se pueden utilizar diferentes tipos de distancia. A lo largo de este trabajo trabajaremos con distancia Euclídea, ya que es la comúnmente más usada para variables cuantitativas de análisis.

3.5 Métodos de Clasificación

Los datos químicos a menudo, surgen de varios grupos o clases que se conocen con anterioridad. Una línea principal dentro de la quimiometría ha sido tratar de resolver problemas de clasificación química, en especial, el reconocimiento automático a partir de datos espectrales moleculares y la asignación del origen de las muestras. Este tipo de aplicaciones han sido conocidas como "reconocimiento de patrones" en química, antes incluso que el término quimiometría sea introducido. Recientemente, los problemas de clasificación han ganado importancia, por ejemplo, para la clasificación de materiales tecnológicos para usando datos de espectroscopía por infrarrojo cercano (Peets, P. et al. 2017), o en aplicaciones médicas (Affonso, C. et al. 2017), y en análisis multivariado de imágenes (Milanez, K. D. T. M. et al. 2017). La identificación de objetos puede ser considerada un caso especial de clasificación, con un solo objeto en cada grupo.

En los problemas de clasificación, es sabido a qué grupos de objetos pertenecen y la hipótesis de trabajo es que las características de los grupos son descriptas por la estructura de datos multivariado de los objetos que incluyen estos grupos. La tarea para el análisis estadístico es resumir esta estructura de datos multivariada apropiadamente para poder establecer reglas para asignar nuevas observaciones para cuales la pertenencia a un grupo determinado es desconocida. Las

reglas usadas para la clasificación deben ser lo más confiables posibles para que el número de objetos mal clasificados sea el menor posible.

El objetivo de la clasificación es establecer reglas –también llamados clasificadores- en la base de objetos que pertenecen a un grupo, los cuales pueden ser usados para predecir la pertenencia de nuevas observaciones, así también como evaluar el desempeño de estas reglas. Hay que tener en cuenta que la clasificación es una técnica supervisada, mientras que la identificación de una estructura de grupos en los datos es una técnica no-supervisada que pertenece al análisis de conglomerados. En un análisis de clústeres la pertenencia a los grupos es desconocida o esa información no está disponible.

Se pueden agrupar los métodos de clasificación según diferentes criterios, en este trabajo se siguió el siguiente:

Métodos Lineales:

- Análisis Discriminante Lineal (LDA).

Métodos No Lineales:

- Árboles de decisión
- Bosques aleatorios (Random Forest)
- Máquinas de soporte vectoriales (SVM)

3.5.1 Métodos Lineales de Clasificación.

3.5.1.1 Análisis Discriminante Lineal o linear discriminant analysis (LDA)

El análisis discriminante lineal es muy similar al de la PCA, mientras que la PCA intenta encontrar los ejes ortogonales de máxima variancia en un conjunto de datos, el

objetivo en LDA es encontrar el subespacio que optimiza la separación de clases. Ambos PCA y LDA son técnicas de transformación lineal que pueden reducir el número de dimensiones en un conjunto de datos, mientras que PCA es un algoritmo no supervisado de datos, LDA es una técnica supervisada de datos. La Fig. 3.6 resume el concepto de LDA para un problema de dos clases (binario). La clase 1 se presentan con cruces y las muestras de la clase 2 se presentan como círculos (Raschka, S. 2015).

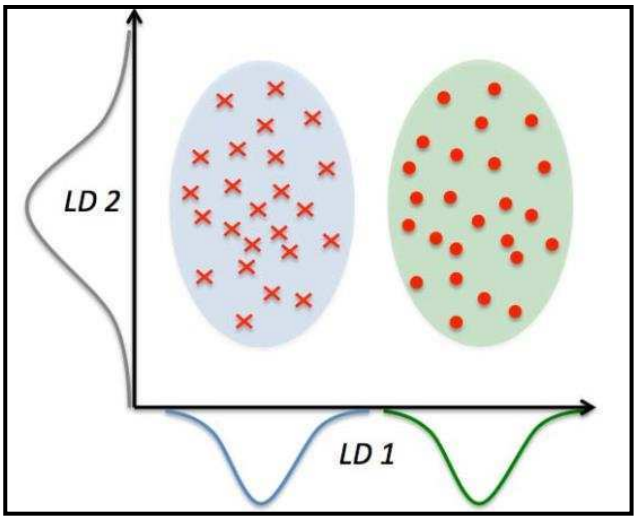


Fig. 3.6 Análisis discriminante lineal de un conjunto de datos binario

Una función discriminante lineal, como se muestra en el eje x (LD1), separará las clases en dos grupos normalmente distribuidos. A pesar de que la segunda función discriminante captura la mayor variancia en el conjunto de datos, fallará para realizar como discriminador ya que no posee información para separar según las clases.

Una asunción del análisis discriminante lineal es que lo datos están normalmente distribuidos. También se asume que las clases tienen matrices de covariancia similares y que las variables son estadísticamente independientes. A pesar de esto si una o más de estas presunciones no se cumplen, el análisis discriminante lineal para una reducción de dimensiones puede funcionar bien (Raschka, S. 2015).

3.5.2 Métodos No Lineales de Clasificación

3.5.2.1 Árboles de decisión o Decision Trees

Los árboles de decisión son poderosos clasificadores, que utilizan una estructura de árbol para modelar las relaciones entre las variables y los posibles resultados. Como se ilustra en la Fig 3.7, la estructura de árbol debe su nombre a que su estructura se asemeja a un árbol. Empieza con un tronco ancho, que se sigue luego de ramas cada vez más y más angostas. De la misma forma, un árbol de decisión clasificatorio usa una estructura de ramas de decisión, que encauza ejemplos en una clase predicha (Lantz, B. 2015).

Para comprender mejor como esto funciona en la práctica, consideremos el siguiente árbol, el cual predice si un trabajo debe ser o no aceptado. Una propuesta laboral para considerar comienza con un nodo raíz, la cual atraviesa luego a través de nodos de decisión que requieren elecciones a ser realizadas basadas en el atributo del trabajo. Estas decisiones dividen los datos a través de ramas que indican los potenciales resultados de esa decisión, mostrados aquí como resultados si o no, a pesar de que en algunos casos puede haber más de dos posibilidades. En el caso de una decisión final, el árbol termina en nodos de hojas (también conocido como nodos terminales) que denotan que la decisión a ser tomada es el resultado de una serie de decisiones. En el caso de un modelo predictivo, los nodos terminales proveen el resultado esperado dado una serie de decisiones previas en el árbol (Lantz, B. 2015). En la Fig. 3.7 se presenta el procedimiento seguido en los árboles de decisión (Lantz, B. 2015).

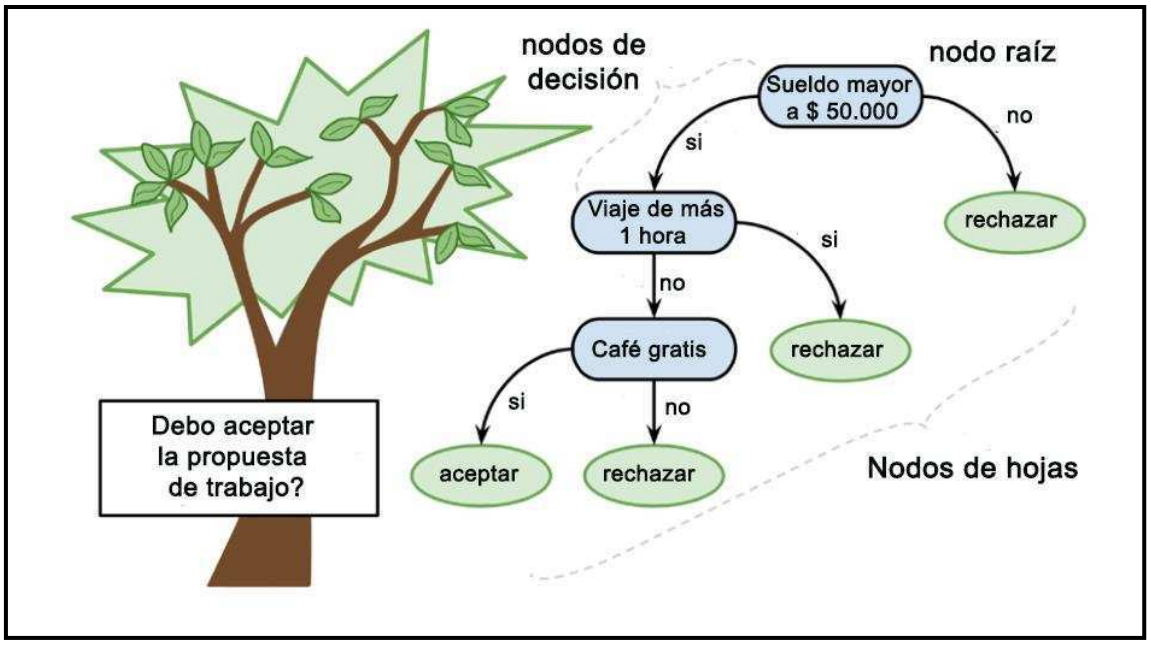


Fig. 3.7 Esquema de procedimiento de un árbol de decisión

El ejemplo anterior muestra el valor de los árboles en el proceso de toma de decisiones, aun así, esto no sugiere que esto termina aquí. De hecho, los árboles de decisión son la técnica de aprendizaje automático más ampliamente usada, y puede ser aplicada para modelar casi cualquier tipo de datos. Aun así, a pesar de su amplia aplicabilidad, vale la pena mencionar algunos casos en los que los árboles de decisión podrían no ser ideales. Un caso sería cuando el conjunto de datos tiene un número de atributos nominales con muchos niveles, o tiene un número mayor de atributos numéricos. Estos casos pueden resultar en un número muy grande de decisiones y en un árbol de estructura muy compleja. Esto también contribuye a la tendencia de los árboles de sobreajustar los datos, aunque esta debilidad del método se puede subsanar ajustando algunos parámetros simples (Lantz, B. 2015).

Los árboles de decisión se construyen bajo un método heurístico de aprendizaje llamado división recursiva, porque divide los datos en grupos, los cuales a su vez se dividen en grupos aún más pequeños, hasta que el proceso se detiene cuando el

algoritmo determina que los datos dentro de los grupos son suficientemente homogéneos, o algún otro criterio de detención se encuentra. Para observar como la división de un grupo de datos puede crear un árbol de decisión, en primer lugar, el nodo raíz representa el conjunto de datos completos, ya que no ha habido división alguna. Luego, el algoritmo debe decidir un atributo para empezar la división, y elige el atributo de mayor poder de predicción dentro de una clase (o grupo) determinado. Los casos luego son divididos en grupos de acuerdo con los valores que sobresalen dentro del atributo seleccionado, y el primer conjunto de ramas del árbol se forma. Trabajando a través de cada rama, el algoritmo continúa dividiendo el conjunto de datos, eligiendo el mejor parámetro cada vez para crear otro nodo de decisión, hasta que un criterio de finalización se encuentra (Lantz, B. 2015). El algoritmo se podría detener en caso de que:

- Todos (o casi todos) los ejemplos en el nodo pertenecen a la misma clase.
- No hay atributos disponibles para seguir separando los casos.
- El árbol ha crecido hasta un tamaño límite predefinido.

En la Tabla 3.1 se resaltan algunas fortalezas y debilidades de los árboles de decisión (Lantz, B. 2015).

Tabla 3.1 Ventajas y Desventajas de los árboles de decisión

Ventajas	Desventajas
Proceso de aprendizaje altamente automático, que trabaja con atributos numéricos o nominales, como así también valores perdidos.	A menudo los arboles de decisión tienen a dividir el set de datos teniendo en cuenta atributos que tienen un gran número de niveles.
Excluye atributos que no son relevantes.	Es fácil sobreajustar el modelo.
Puede ser en set de datos pequeños como grandes.	Pequeño cambio en el conjunto de datos de entrenamiento puede resultar en cambios en la decisión lógica de la división
Resulta en un modelo que puede ser interpretado sin un conocimiento matemático profundo.	Los árboles grandes pueden ser difíciles para interpretar y las decisiones pueden parecen contra-intuitivas.

Más eficiente que otros modelos más complejos.

3.5.2.2 Bosques Aleatorios o Random Forests (RFs)

La creación de un árbol de decisión provee un modelo simple, pero es a menudo, muy simple o muy específico. Después de muchos años de experiencia, se ha hecho claro que varios modelos trabajando juntos es mejor que un solo modelo realizando todo el trabajo. Ahora se ha hecho familiar la idea de combinar muchos modelos, como los árboles de decisión, en un solo modelo de ensamble, para crear un bosque de árboles. La técnica de Random Forests es una técnica que funciona bajo el del ensamble de algoritmos de bajo aprendizaje (árboles de decisión, en este caso) para mejorar el porcentaje de acierto de la técnica global.

El algoritmo de bosques aleatorios tiende a producir modelos muy precisos porque el conjunto reduce la inestabilidad que se puede observar cuando construimos arboles de decisión. Esto puede observarse cuando se remueve un número pequeño de observaciones desde el conjunto de datos de prueba, para ver cuánto cambian los resultados en los árboles de decisión. Los algoritmos de ensamble o conjunto, como los bosques aleatorios, tienden a ser más robustos a los cambios en el conjunto de datos. Se puede decir que es un método robusto al ruido, esto implica que pequeños cambios en el conjunto de datos de entrenamiento, tendrán un impacto bajo o ínfimo si es que lo tuviesen. Los bosques aleatorios es una técnica muy eficaz en casos de clasificación no lineal como las máquinas de soporte vectoriales o las redes neuronales.

Mediante la creación de árboles de decisión en su máxima profundidad, como el algoritmo de bosques aleatorios, se obtiene un modelo que presenta menos error. Cada árbol de decisión va a sobreajustar los datos, pero mediante muchos árboles usando diferentes variables, el ajuste del modelo a los datos resulta diferente. La aleatoriedad introducida por el método se presenta tanto en la selección de observaciones y variables. Es esta aleatoriedad que permite que el método sea robusto al ruido, a los datos extremos y al sobreajustado, cuando se lo compara con un solo árbol de decisión. La aleatoriedad también tiene beneficios en cuanto al costo computacional. La creación de un solo árbol de decisión implica la elección de un grupo de observaciones disponibles en los datos de entrenamiento. También, en cada nodo durante el proceso de la creación de los bosques aleatorios, solo una pequeña fracción de las variables disponibles son consideradas cuando se determina la mejor separación del conjunto de datos. Esto reduce los requerimientos en cuanto al costo computacional. En resumen, el modelo de los bosques aleatorios es una buena elección por varios motivos. A menudo, muy poco preprocesamiento de los datos es necesario realizar, los datos no necesitan ser normalizados y es bastante robusta a los datos extremos. La necesidad de selección de variables se evita ya que el algoritmo lo realiza durante el análisis. Como muchos árboles se construyen a dos niveles de aleatoriedad (observaciones y variables), cada árbol es un modelo independiente y el modelo resultante tiende a no sobreajustar el set de datos de entrenamiento (Williams, G. 2011).

En la Tabla 3.2 se resumen ventajas y desventajas de método Random Forest (Lantz, B. 2015).

Tabla 3.2 Ventajas y desventajas de Random Forest

Ventajas	Desventajas
Es un modelo que funciona bien en la mayoría de los problemas (clasificación y regresión). Puede lidiar con datos perdidos (missing data) o con datos que presentan ruido Selecciona sólo las variables más importantes. Puede ser utilizado en conjunto de datos que son extremadamente grandes o con un gran conjunto de variables.	A diferencia de los árboles de decisión este modelo no es fácilmente interpretable. Puede llevar un tiempo hasta lograr optimizar los hiperparámetros.

3.5.2.3 Máquinas de Soporte Vectorial o Support Vector Machines (SVMs)

El término "support vector machine" se refiere a una técnica de aprendizaje automático que puede ser usada para clasificación y regresión. En el contexto de la clasificación, se producen límites lineales entre grupos de objetos en un espacio transformado de las variables *m* que usualmente es de una dimensión mayor que el espacio x original. La idea de crear un espacio dimensional más grande es crear grupos que puedan separarse en ese nuevo espacio. Entonces, en el espacio multidimensional, los límites de las clases se construyen de tal modo de maximizar la separación entre los grupos. Una comparación de SVMs con otros métodos de clasificación y regresión muestra que ellos tienen un gran desempeño, a pesar de que otros métodos demuestran ser muy competitivos (Lantz, B. 2015).

Mientras la mayoría de los métodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento (error empírico), el sesgo inductivo asociado a las SVMs radica en la minimización del denominado riesgo estructural. La idea es seleccionar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de

definir el hiperplano, sólo se consideran los ejemplos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes. Estos ejemplos reciben el nombre de vectores soporte. Desde un punto de vista práctico, el hiperplano separador de margen máximo ha demostrado tener una buena capacidad de generalización, evitando en gran medida el problema del sobreajuste a los ejemplos de entrenamiento (Lantz, B. 2015).

Una máquina de soporte vectorial puede ser entendida como una superficie que crea un límite entre los puntos de los datos graficados en un espacio multidimensional que representan ejemplos y los valores de sus atributos. El objetivo de la técnica SVM es crear un límite llamado hiperplano, que divide el espacio para crear divisiones apenas homogéneas en cada lado. De esta forma, el aprendizaje mediante SVMs combina aspectos del aprendizaje del tipo k-N vecinos más cercanos (kNN) y de regresión lineal. La combinación es extremadamente poderosa, permitiendo al algoritmo SVM modelar relaciones altamente complejas.

Las máquinas de soporte vectorial pueden ser adaptadas para prácticamente cualquier tipo de aprendizaje, incluyendo clasificación y predicción numérica. Gran parte del éxito del algoritmo ha sido en el campo del reconocimiento de patrones. El método SVMs es fácilmente entendible cuando es utilizado para clasificación binaria, que es la forma en la que el método ha sido tradicionalmente aplicado. (Lantz, B. 2015).

3.5.2.3.1 Clasificación mediante hiperplanos

Como se dijo anteriormente, el método SVMs utiliza un límite llamado hiperplano para dividir el set de datos en grupos de clases similares. Por ejemplo, la

siguiente figura muestra un hiperplano que separa grupos de círculos y cuadrados en dos y tres dimensiones. Porque los círculos y los cuadrados pueden ser separados por una línea recta o un hiperplano, se dice que son linealmente separables. En principio consideraremos este caso, pero el algoritmo SVMs puede ser extendido a problemas donde los puntos no son linealmente separables (Lantz, B. 2015). En la Fig. 3.8 se muestra la línea y el hiperplano capaces de separar datos bidimensionales y tridimensionales (Lantz, B. 2015).

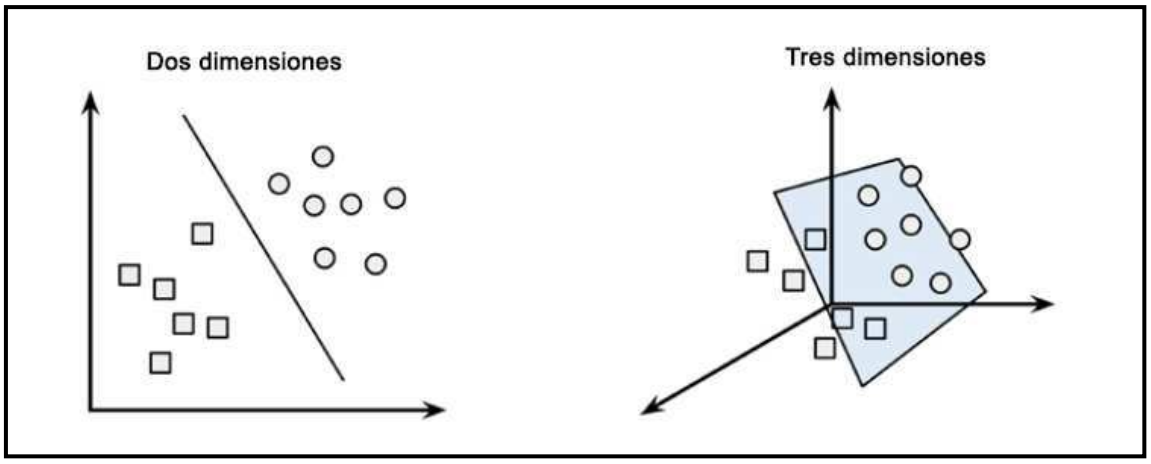


Fig. 3.8 Línea e hiperplano que separan un conjunto de datos mediante el algoritmo SVMs.

En dos dimensiones, la tarea de las máquinas de soporte vectoriales es identificar una línea que separa los dos grupos. Como se presenta en la Fig. 3.9, hay más de una opción para separar entre el grupo de círculos y cuadrados. Las tres posibilidades se pueden llamar a, b y c (Lantz, B. 2015).

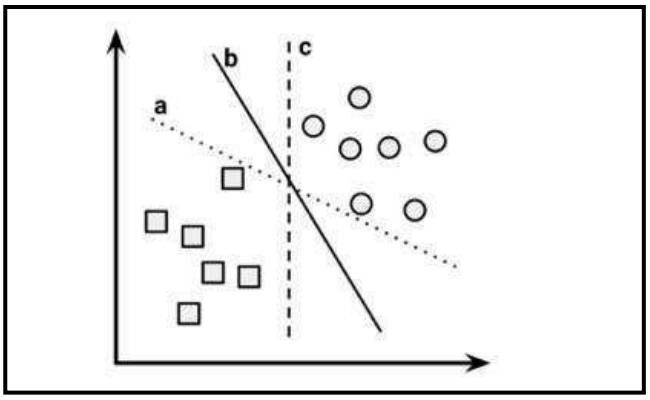


Fig. 3.9 Tres posibilidades de líneas que pueden separar un conjunto de datos linealmente separable.

¿Cuál de estas tres líneas elegirá el algoritmo? La respuesta a esta pregunta implica una búsqueda del hiperplano de máximo margen que crea la mayor separación entre las dos clases. A pesar de que las tres líneas separan los círculos de los cuadrados, el algoritmo elige la recta que conlleva la mayor separación que generaliza lo mejor para el conjunto de datos. El máximo margen mejorará la chance de que, a pesar del ruido aleatorio, los puntos se quedaran en el lado correcto del límite de decisión (Lantz, B. 2015).

Los vectores de soporte son los puntos de cada clase que están más cerca al hiperplano de máximo margen; cada clase debe tener al menos un vector de soporte, pero es posible que exista más de uno. Utilizando estos vectores de soporte es posible definir el hiperplano de máximo margen. Esta es una característica clave del algoritmo SVMs; el cual provee un modelo de clasificación, aún si el número de atributos es extremadamente largo (Lantz, B. 2015). En la Fig. 3.10 se observa el máximo margen para una separación de datos mediante SVM (Lantz, B. 2015).

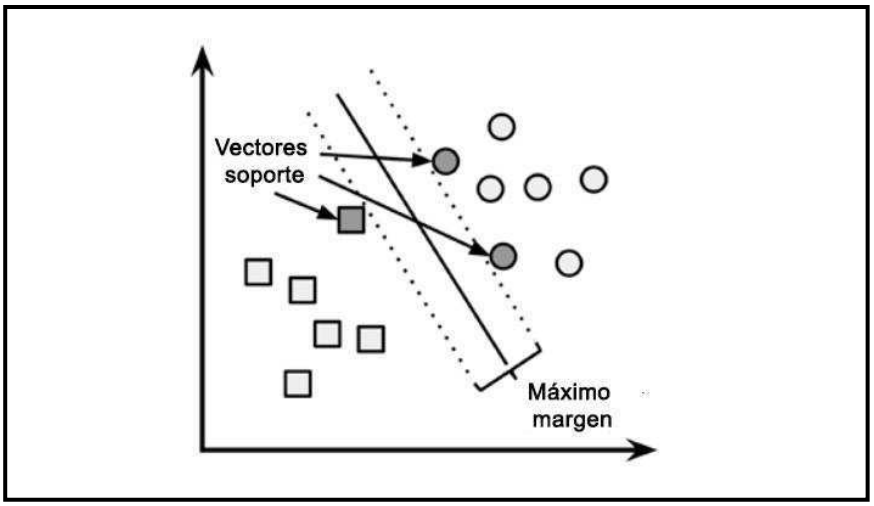


Fig. 3.10 Esquema de los vectores de soporte que permite trazar la recta del algoritmo SVMs

3.5.2.3.2 Los datos linealmente separables

Es más fácil entender cómo encontrar el máximo margen bajo la suposición que los grupos son linealmente separables. En este caso, el máximo margen del hiperplano (MMH) se obtiene desde los límites de los dos grupos del set de datos. Estos límites externos se llaman cascos convexos. Es el MMH entonces el bisector perpendicular de la línea más corta entre los cascos convexos. Algoritmos computacionales que utilizan esta técnica conocida como optimización cuadrática son capaces de encontrar el máximo margen de este modo (Lantz, B. 2015). En la Fig. 3.11 se observa el MMH entre dos cascos convexos en una separación mediante SVM (Lantz, B. 2015).

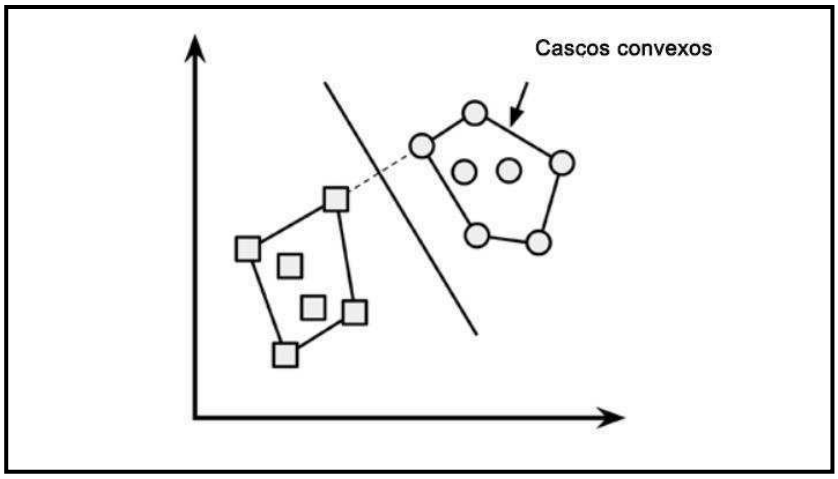


Fig. 3.11 Máximo margen del hiperplano entre dos cascos convexos.

Un enfoque alternativo tiene en cuenta una búsqueda a través del espacio de cada posible hiperplano para encontrar un conjunto de dos planos paralelos que dividen los puntos en grupos homogéneos tanto como sea posible. Para entender el proceso de búsqueda, necesitaremos definir exactamente lo se entiende por hiperplano. En un espacio n-dimensional, la siguiente ecuación se utiliza:

$$\vec{w} \,.\, \vec{x} + b = 0$$

Mientras que w es un vector de n componentes $\{w_1, w_2, \ldots, w_n\}$ y b es el bias o sesgo estadístico.

$$\vec{w} \,.\, \vec{x} + b \geq +1$$

$$\vec{w} \,.\, \vec{x} + b \leq +1$$

También se necesita que estos hiperplanos estén especificados para que todos los puntos de un grupo estén debajo de uno de los hiperplanos y los otros puntos del otro grupo estén por arriba del segundo hiperplano. Esto es posible siempre y cuando los grupos sean linealmente separables (Lantz, B. 2015).

3.5.2.3.3 Los datos no separables linealmente

La solución a este problema son las variables latentes, que crean un margen débil que permite que algunos puntos caigan en el lado incorrecto del margen. La Fig 3.12 ilustra dos puntos que están del lado incorrecto de la línea con los correspondientes términos débiles (Lantz, B. 2015). Cuando el conjunto de datos no es separable mediante un hiperplano, se hace necesario relajar las restricciones de la ecuación de optimización. Esto se logra agregando un término más en la ecuación de optimización, con ε_i junto a el parámetro de coste, C (Lantz, B. 2015).

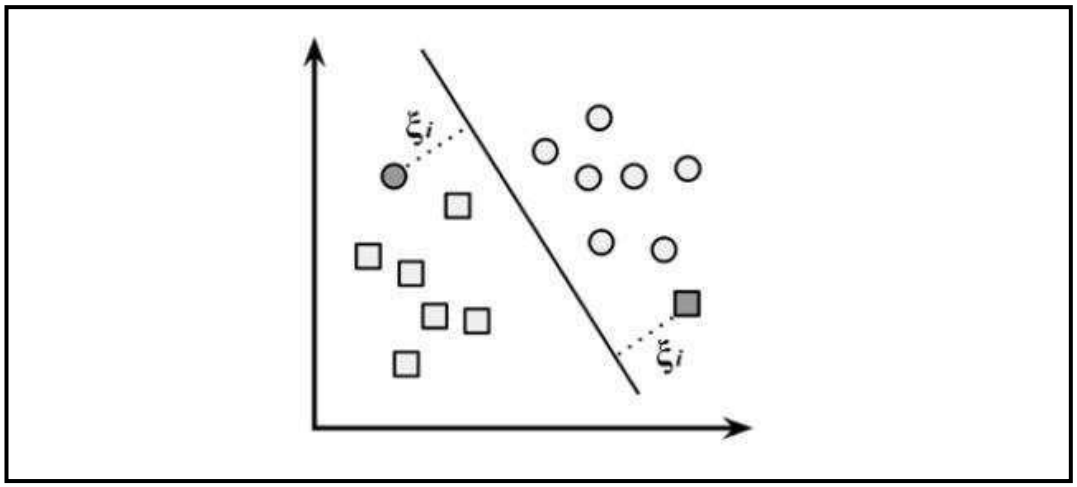


Fig. 3.12 Conjunto de datos que no es separable mediante un hiperplano.

Un valor de coste, denotado por C, se aplica a todos los puntos que violan las restricciones impuestas y en vez de buscar el máximo margen de separación, el algoritmo intenta minimizar el valor de coste, C. Entonces podemos replantear el problema de optimización como

$$\min \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^{n} \varepsilon_i$$

$$s.t.\, y_i\,(\vec{w}\,.\vec{x_i} - b) \geq 1 - \varepsilon_i, \forall\, \vec{x_i}\,, \varepsilon_i \geq 0$$

Lo importante es entender la adición del parámetro, valor de coste, C. Modificando este valor, se ajusta la penalidad, es decir que el caso mal clasificado no vuelva a caer en lado incorrecto del hiperplano. Cuanto mayor sea el valor de coste, C, más intentará la optimización lograr una separación 100%. En el otro extremo, un bajo valor de C pondrá énfasis en un margen más laxo. Es importante establecer un equilibrio entre estos dos, para establecer un modelo que pueda ser generalizable a datos futuros (Lantz, B. 2015). En la Fig. 3.13 se ilustra las consecuencias de la variación (aumento) de valor de C en el límite de decisión de SVM.

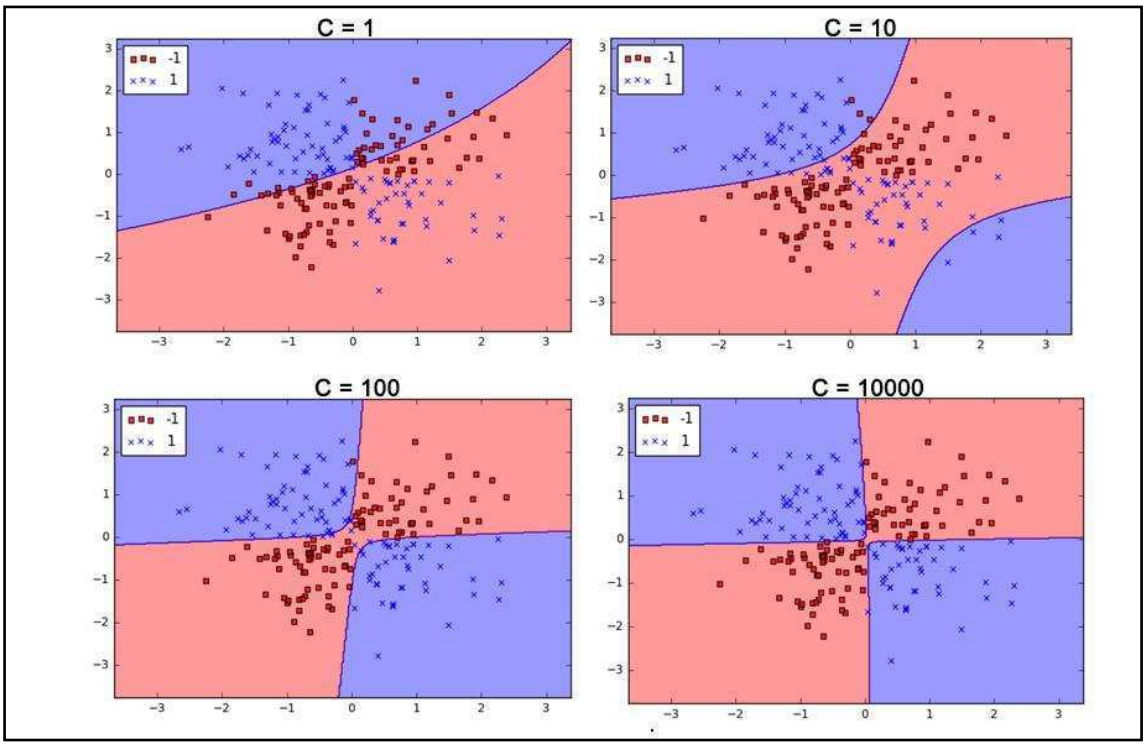


Fig 3.13 Consecuencia del aumento de valor de C en los límites de decisión de SVM.

3.5.2.3.4 El uso de kernels para casos no linealmente separables

En muchas aplicaciones reales, las relaciones entre variables son no-lineales. Una característica fundamental del algoritmo SVMs es su habilidad para mapear el problema en dimensiones más grandes usando un proceso conocido como el truco del

kernel. Empleando esta técnica una relación no-lineal puede aparecer como lineal. Esto puede ilustrarse mediante la Fig. 3.14 (Lantz, B. 2015)

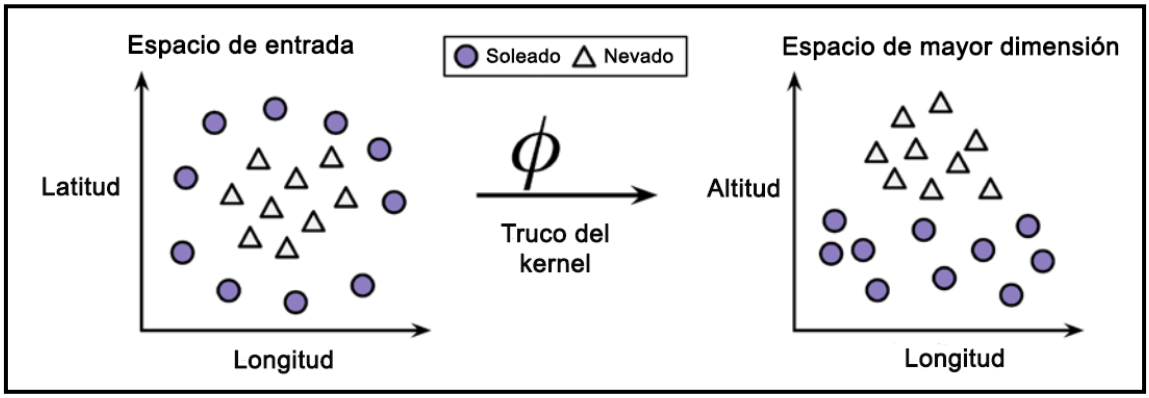


Fig. 3.14 Truco del kernel en SVM

En la primera imagen se observa un grupo de observaciones que muestra una disposición no lineal en un gráfico de dos dimensiones, cuando se grafican la longitud y la latitud. Sin embargo, en la figura de la derecha, luego de aplicar el truco del kernel, se pueden observar los datos con respecto a otra variable que es la altitud. Se observa que ahora las clases son linealmente separables, porque se observan los datos desde una nueva perspectiva (Lantz, B. 2015).

La técnica de SVM con kernels no lineales agrega dimensiones adicionales a los datos para crear una separación en este sentido. En esencia, el truco del kernel implica un proceso de nuevos atributos que expresan la relación matemática entre los atributos medidos (Lantz, B. 2015).

En principio, la altitud puede ser expresada matemáticamente como una interacción entre la latitud y la longitud. Esto permite a la técnica de SVM aprender nuevos conceptos que no fueron medidos explícitamente en el conjunto de datos originales (Lantz, B. 2015). En la Tabla 3.3 se resumen las principales características de este método (Lantz, B. 2015).

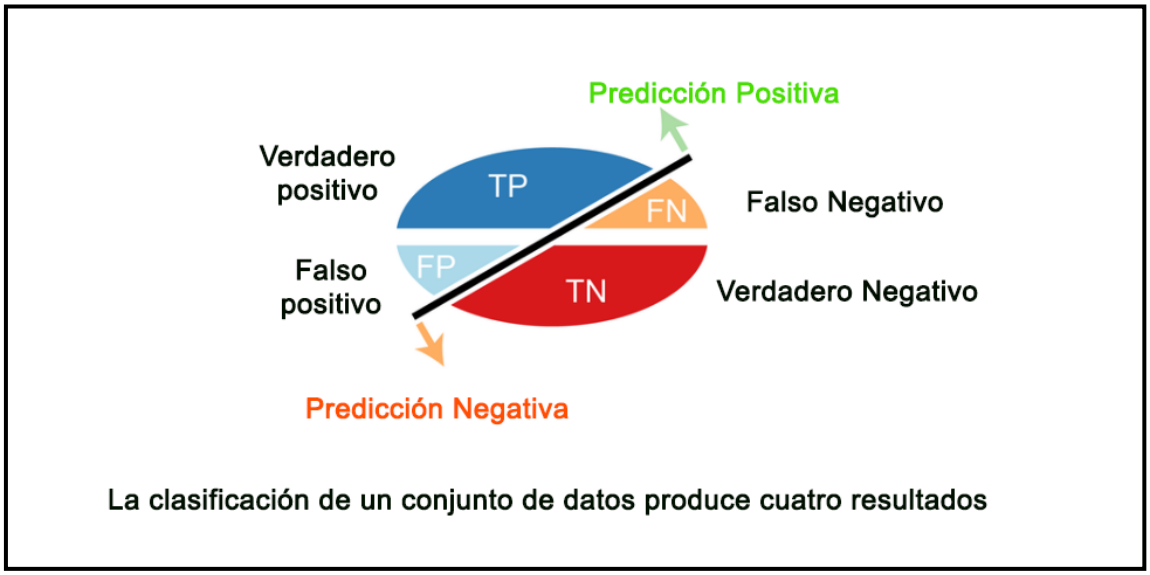
Tabla 3.3 Ventajas y desventajas de la técnica SVM

Ventajas	Desventajas
Puede ser utilizada para clasificación como para predicción de valores numéricos.	Encontrar el mejor modelo requiere optimizar los parámetros C y σ.
No está influenciada por el ruido de los datos.	El entrenamiento puede ser lento dependiendo de la cantidad de atributos (o columnas) que tenga el conjunto de datos.
Es más fácil de optimizar que las redes neuronales (ANNs).	Es un modelo de caja negra que es difícil de interpretar.
Alto porcentaje de acierto (exactitud) dentro de las técnicas de minería de datos.	Ideal para matrices de datos pequeñas (alrededor de 100 muestras)

3.6 Evaluación de Modelos Quimiométricos

3.6.1 Matriz de Confusión

La matriz de confusión para un caso binario contiene cuatro valores carácterísticos: verdaderos positivos (VP), falso positivos (FP), falso negativos (FN), y verdaderos negativos (VN). Los casos verdaderos positivos (VP) y los verdaderos negativos (VN) son los casos que han sido clasificados correctamente mientras que los casos de falso negativo (FN) y falso positivo (FP) son casos clasificados erróneamente. Estos cuatro parámetros se presentan en la Fig. 3.14 (Takaya, S. and Rehmsmeier, M. 2015).



La clasificación de un conjunto de datos produce cuatro resultados

Fig. 3.15 Resultados posibles de la matriz de confusión

3.6.1.1 Exactitud

La exactitud es la suma de los casos positivos y los casos negativos correctamente clasificados sobre el total de las muestras utilizadas en el proceso de clasificación. En la Fig. 3.16 se ilustra el concepto de exactitud (Takaya, S. and Rehmsmeier, M. 2015).

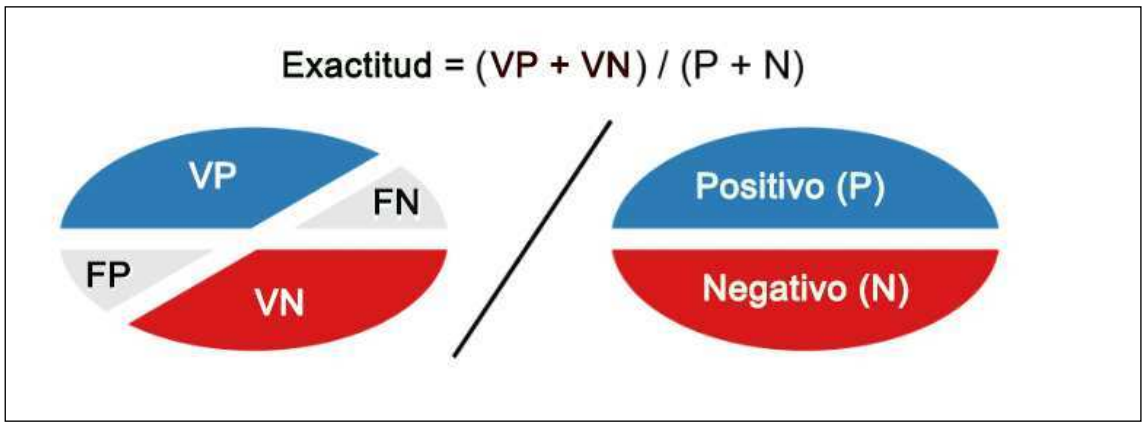


Fig. 3.16 Exactitud de un clasificador

3.6.1.2 Sensibilidad y Especificidad

La tasa de verdaderos positivos (true positive rate) es también llamada sensibilidad de un clasificador. Este término se origina en el campo médico, en el que la métrica típicamente utilizada para estudiar la efectividad de una prueba clínica en detectar una enfermedad. El proceso de la evaluación en el contexto de la detección de una enfermedad es equivalente a investigar cuán sensible es un test en la presencia de la enfermedad. ¿Esta es la forma en la que muchas instancias positivas (casos de enfermedad) puede ser detectada exitosamente? La métrica complementaria a esto se enfocaría en la proporción de instancias negativas que son detectadas. Esta métrica es la especificidad de un algoritmo de aprendizaje. Por lo tanto, la especificidad es la tasa de casos negativos de un clasificador binario. Esto es, mientras la sensibilidad tiene en

cuenta los casos positivos mientras que la misma cantidad, cuando se la mide en los casos negativos, se refiere a la especificidad (Japkowicz, N. and Shah, M. 2011). En las Fig. 3.17 y 3.18 se ilustran los conceptos de sensibilidad y especificidad (Takaya, S. and Rehmsmeier, M. 2015).

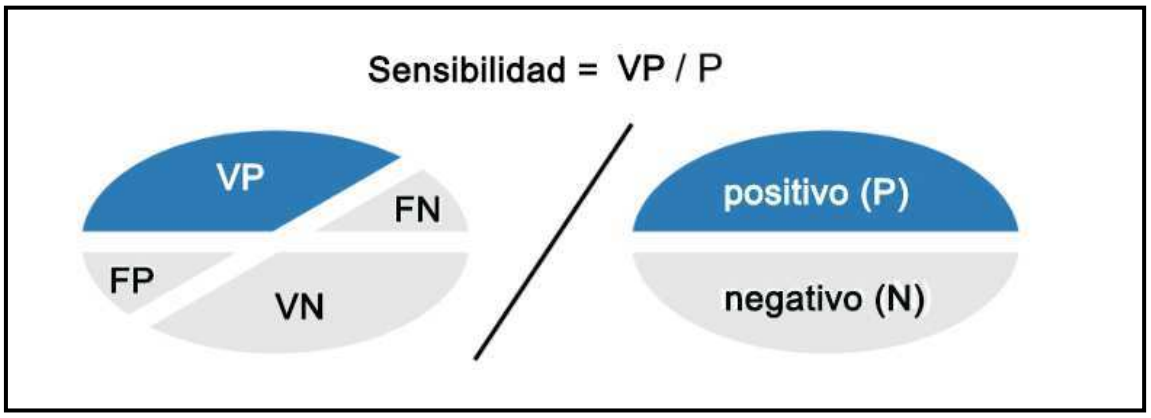


Fig. 3.17 Sensibilidad de un clasificador

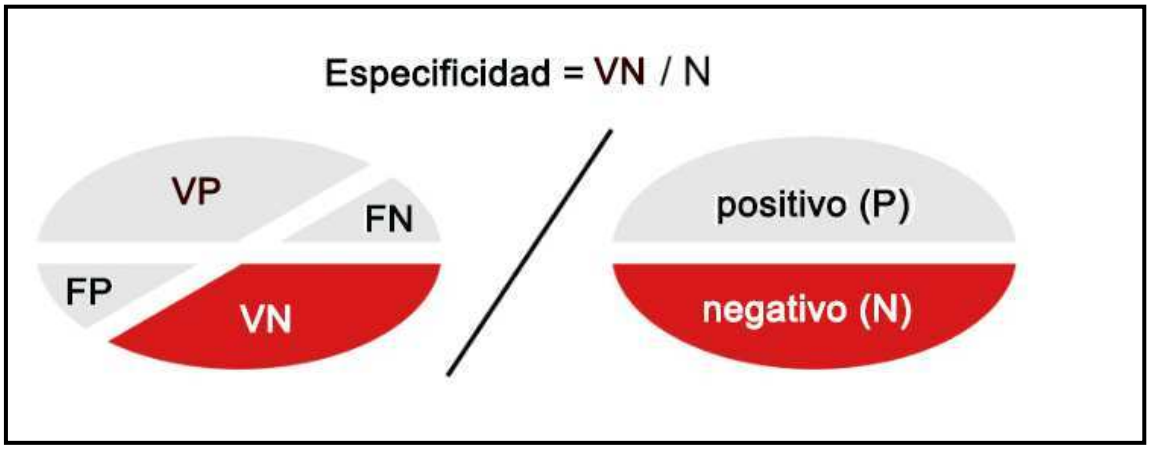


Fig. 3.18 Especificidad de un clasificador

3.6.1.3 Precisión

Otro aspecto para la evaluación es la pregunta de qué proporción de muestras que verdaderamente pertenecen a una determinada clase de todos los ejemplos asignados (o clasificados) como tal. El valor de los positivos predichos (PPV) mide esto al evaluar un clasificador particular. Como esta métrica evalúa cuan *preciso* es el algoritmo para identificar los casos de una clase dada. El término precisión es más

común en el campo de la búsqueda y recuperación de la información (information retrieval) (Japkowicz, N. and Shah, M. 2011). En la Fig. 3.19 se ilustra el concepto de precisión (Takaya, S. and Rehmsmeier, M. 2015).

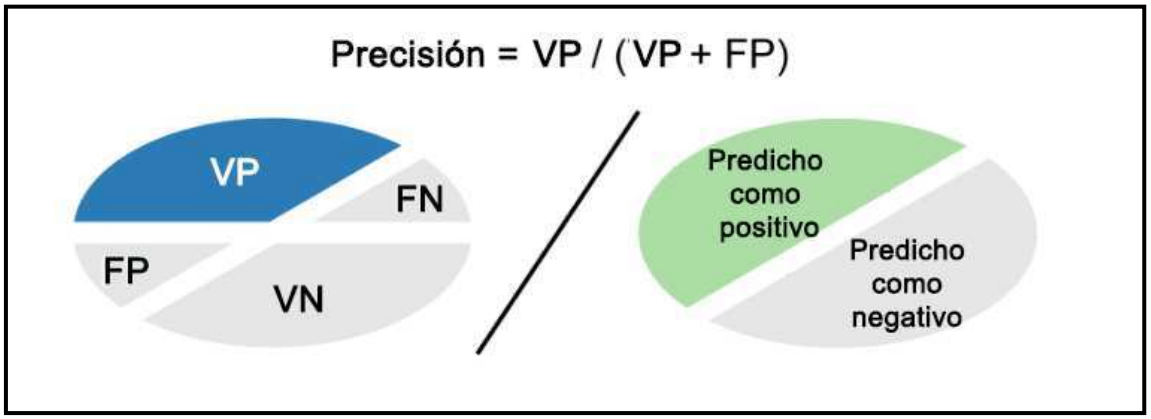


Fig. 3.19 Precisión de un clasificador

A continuación, se analizan métodos gráficos para evaluar el desempeño de un clasificador.

3.6.2 Métodos gráficos de evaluación de modelos de clasificación

3.6.2.1 Curvas ROC

En el contexto de los algoritmos de aprendizaje, las curvas ROC han sido usadas de una variedad de formas. Las curvas ROC son una poderosa herramienta gráfica para visualizar el desempeño de un algoritmo de aprendizaje a través de una variación en el criterio de decisión, comúnmente en una clasificación binaria. Las curvas ROC se han utilizado no sólo para estudiar el comportamiento de un algoritmo sino también para identificar regiones óptimas de desempeño, realizar la selección de modelos, y lo más importante para comparar y evaluar algoritmos de aprendizaje automático.

Una curva ROC es un gráfico que tiene el eje horizontal (eje x) que muestra la tasa de falsos positivos (TFP) y el eje vertical (eje y) muestra la tasa de verdaderos positivos (TVP) de un clasificador. Como hemos dicho anteriormente, la tasa de verdaderos positivos (TVP) no es más que la sensibilidad, y la tasa de falsos positivos (TFP) no es más que 1 – TVN, siendo TVN equivalente a la especificidad. En este sentido, un análisis de la curva ROC estudia la relación entre la sensibilidad y la especificidad de un clasificador. En la Fig. 3.20 se ilustran estos conceptos (Takaya, S. and Rehmsmeier, M. 2015).

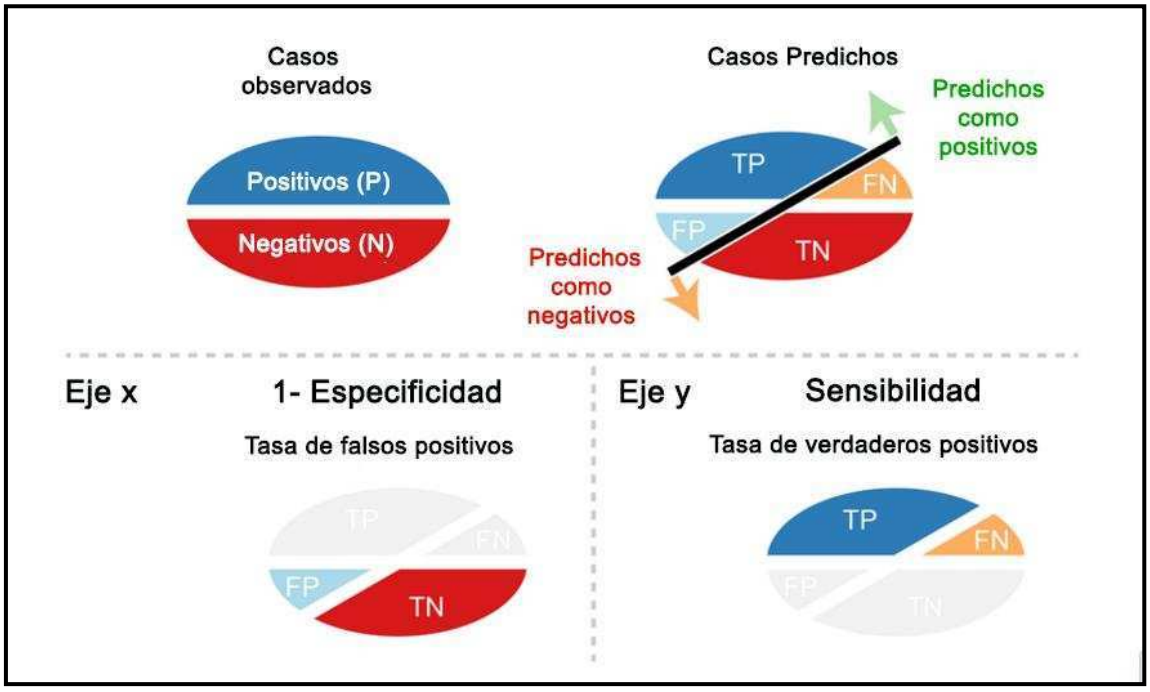


Fig. 3.20 Métricas implicadas en una curva ROC

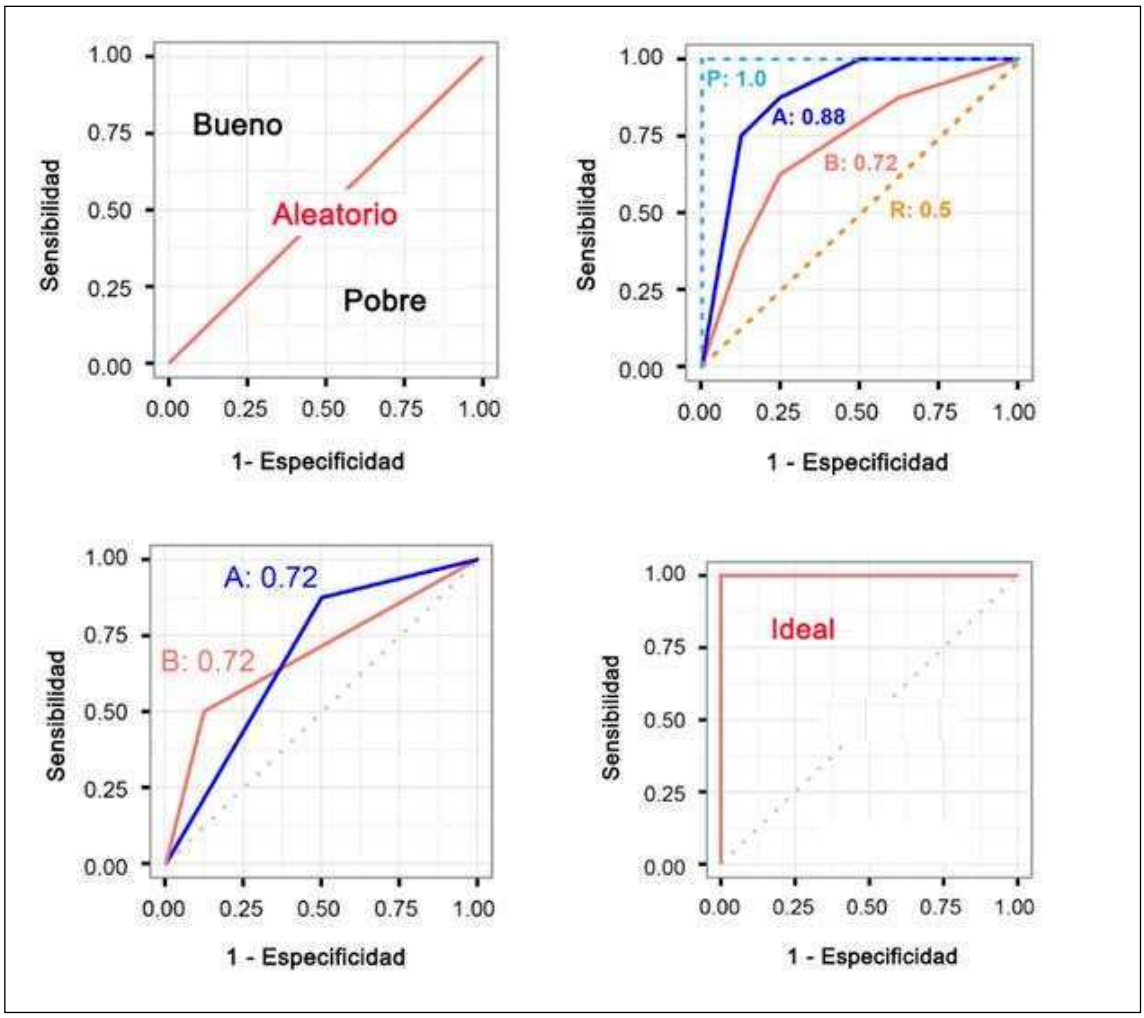


Fig 3.21 Diferentes situaciones en las curvas ROC

En la Fig 3.21 se observan diferentes situaciones que se presentan en el análisis de una curva ROC de un clasificador binario.

- El desempeño de un clasificador aleatorio, mostrando dos zonas: una de clasificación buena a óptima y otra, de clasificación pobre.
- Cuanto mayor es el área bajo la curva es mejor la clasificación binaria.
- Dos valores numéricos iguales de área bajo la curva no implican el mismo gráfico de curva ROC.
- El clasificador ideal nos muestra un área igual a 1.

3.6.2.2 *Curvas Precisión-Exactitud, Precision-Recall o Curvas PR*

Las curvas de precisión-exactitud, a veces abreviado como curvas PR, son similares a las curvas ROC en el sentido de que exploran la relación entre los casos positivos bien clasificados y el número de casos negativos mal clasificados. Como el nombre lo indica, las curvas PR grafican la precisión de un clasificador como función de su sensibilidad. En otras palabras, mide la cantidad de precisión que puede obtenerse cuando varios grados de sensibilidad son considerados. La diferencia con respecto a una curva ROC es que tiene una pendiente negativa. Esto se debe a que la precisión decrece a medida que la sensibilidad aumenta. Las curvas PR pueden ser más útiles que las curvas ROC cuando se está frente a datos que presentan desbalance de clases, es decir, cuando la proporción de una clase es mucho mayor que otra (Japkowicz, N. and Shah, M. 2011). En la Fig. 3.22 se presenta esquemáticamente las métricas implicadas en las curvas PR (Takaya, S. and Rehmsmeier, M. 2015).

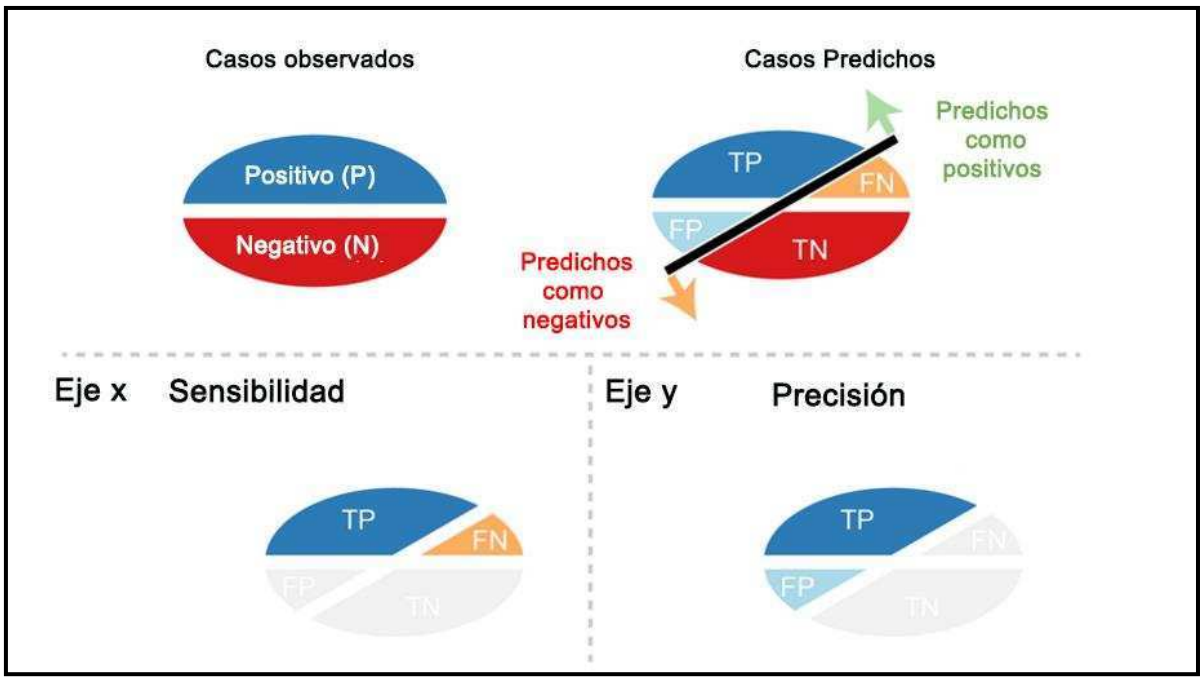


Fig. 3.22 Métricas implicadas en una curva PR

En la Fig 3.23 se observan varias situaciones para el análisis de las curva PR.

- Se observa la línea de clasificación mediante un clasificador aleatorio, que delimita dos zonas: una zona de clasificación buena, y otra de una clasificación pobre.
- Al mayor ser el área bajo una curva PR, más se acerca al óptimo de 1.
- Un mismo valor de área bajo la curva ROC, no implica que el área bajo la curva PR sea igual numéricamente.
- El área bajo la curva para un clasificador ideal es la misma, y tiene un valor numérico de 1. El final de la línea llega hasta donde se considera está el valor de P / (P + N), siendo P la suma de verdaderos positivos y falsos negativos, y N, la suma de verdaderos negativos, y falsos positivos.

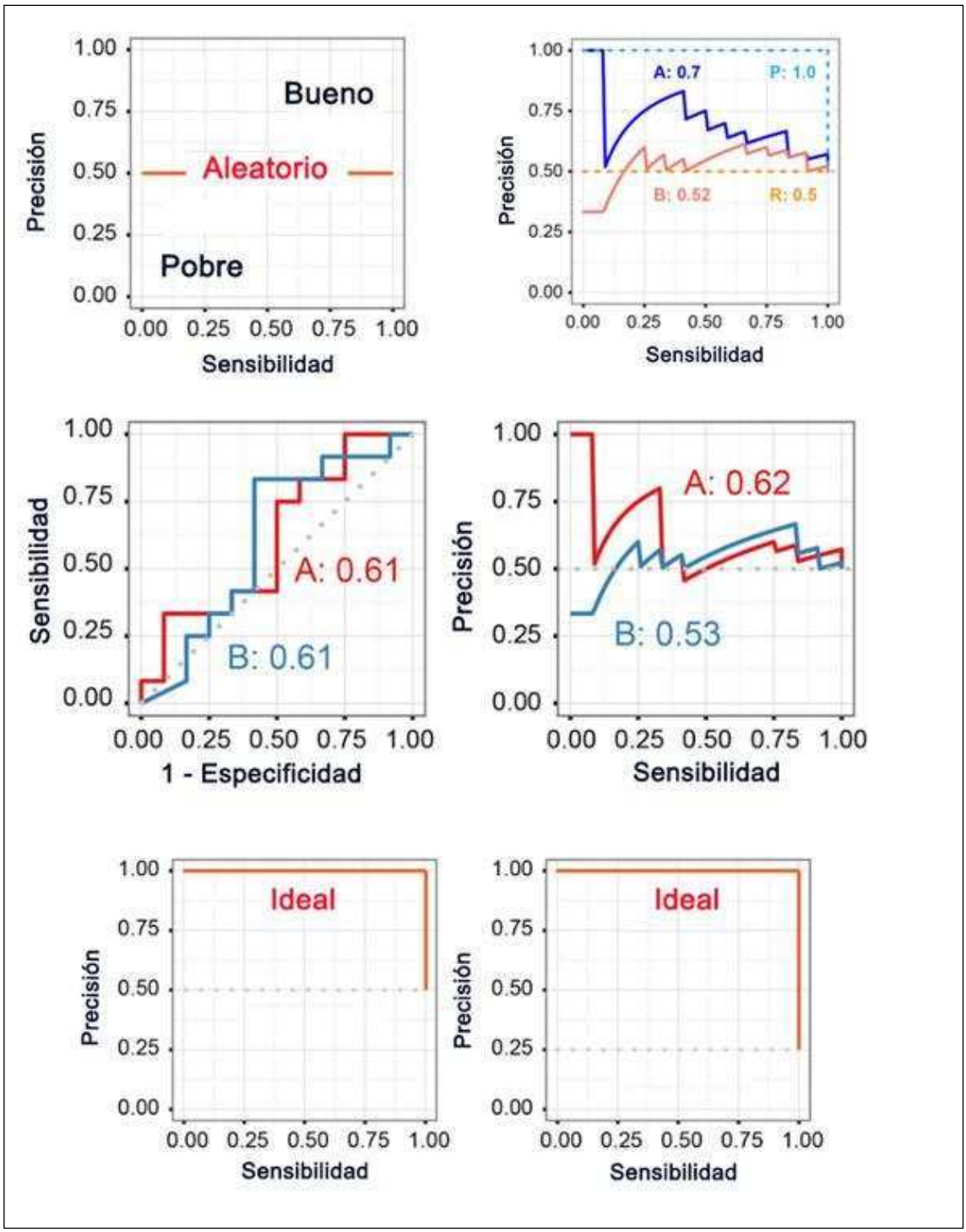


Fig 3.24 Diferentes situaciones en una curva PR

3.7 Referencias Bibliográficas

Affonso C, Rossi ALD, Vieira FHA, de Carvalho ACPdLF. (2017) Deep learning for biological image classification. Expert Systems with Applications.85:114-122.

Brereton RG. (2015) Pattern recognition in chemometrics. Chemom Intell Lab Syst.149, Part B:90-96.

Harrington P. (2011) Machine Learning in Action: Manning Publications Company.

Hastie T, Tibshirani R, Friedman J. (2013) The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Springer.

Japkowicz N, Shah M. (2011) Evaluating Learning Algorithms: A Classification Perspective: Cambridge University Press.

Kumar N, Bansal A, Sarma GS, Rawal RK. (2014) Chemometrics tools used in analytical chemistry: An overview. Talanta.123:186-199.

Lantz B. (2015) Machine Learning with R: Packt Publishing Ltd.

Milanez KDTM, Nóbrega TCA, Nascimento DS, Insausti M, Pontes MJC. (2017) Transfer of multivariate classification models applied to digital images and fluorescence spectroscopy data. Microchemical Journal.133:669-675.

Peets P, Leito I, Pelt J, Vahur S. (2017) Identification and classification of textile fibres using ATR-FT-IR spectroscopy with chemometric methods. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy.173:175-181.

Raschka S. (2015) Python Machine Learning: Packt Publishing Ltd.

Takaya S, Rehmsmeier M. 2015. Classifier evaluation with imbalanced datasets. In: https://classeval.wordpress.com/: Wordpress.

Varmuza K, Filzmoser P. (2009) Introduction to Multivariate Statistical Analysis in Chemometrics: CRC Press.

Williams G. (2011) Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery: Springer Science & Business Media

SECCIÓN II:
PARTE
EXPERIMENTAL

4) Capítulo IV

4.1 Muestreo

La zona campos abarca principalmente la Provincia de Corrientes (27º y 30º lat. S, 53 y 59º long. W) con una superficie de 8.920.000 ha. El clima es subtropical, la temperatura media anual varía de 21º C en el N a 19º C en el S. El período libre de heladas abarca desde octubre hasta abril. Las precipitaciones son de 1.500 mm en el NE, descendiendo a 1.000 mm en el SW. La época más lluviosa es el otoño y la más seca el invierno, en el verano hay déficit hídrico, las precipitaciones son menores que la evapotranspiración.

Identificada las series de suelo, en cada sitio (Fig. 4.1), se realizó colectas de material vegetal. Una vez identificadas directamente en campo (con acompañamiento del profesional de apoyo) las especies se procedió a realizar una colección en el terreno de muestras que fueron preparadas de acuerdo con procedimientos estándares de herborización. Estos procedimientos incluyeron la colección de uno a tres ejemplares por especie, tomando la muestra más completa posible. Los ejemplares recolectados fueron prensados en el campo y etiquetados siguiendo la metodología taxonómica convencional y se enviaron al Herbario del Instituto de Botánica del Nordeste (IBONE) en Corrientes, Capital, donde fueron identificadas o confirmada su identidad.

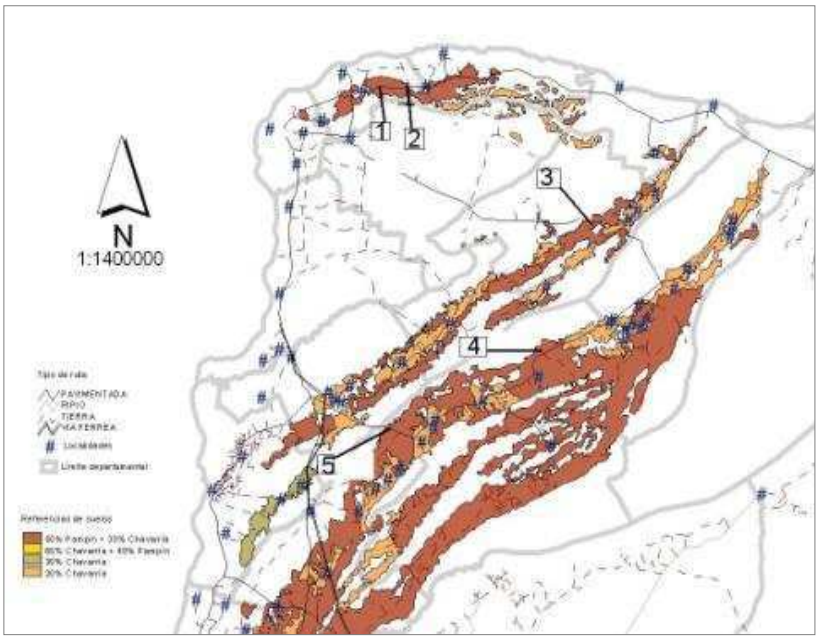


Fig. 4.1. Los puntos 1 al 5 representan los sitios propuesto para el muestreo de suelo y vegetación.

El experimento se condujo en la región occidental de la provincia de Corrientes, República Argentina, en los Departamentos San Cosme, General Paz, San Miguel, Concepción y San Roque. El área de referencia se encuentra situada entre los paralelos 27°20’ y 28°21’ de latitud Sur, los meridianos 57°12’ y 58°41’ long. Oeste, entre 60 y 72 m.s.n.m. De acuerdo con el mapa de suelos de la provincia de Corrientes (Escobar, et al, 1996), pertenece a la Región de Albardón y planicie subcóncava del Río Paraná y afluentes y a la Región de las Lomadas arenosas, Planicies y depresiones.

Se identificaron en el área de referencia las Serie de suelo Chavarria y Pampín, según el Mapa de Suelos de la provincia de Corrientes (Escobar et al, 1996) (Tabla 4.1)

Tabla 4.1 Clasificación taxonómica de las series de Suelo

Orden	Suborden	Gran Grupo	Sub grupo	Familia	Serie
ENTISOLES	ACUENTES	PSAMACUENTES	SPODICOS	Arenosa	Chavarria
			TIPICOS	Arenosa	Pampín

La Serie de suelo Chavarria constituye una de las series de mayor distribución y superficie dentro de la provincia de Corrientes. El relieve es normal y se ubica en planicies arenosas pardo amarillentas, en posición de media loma a media loma baja, con pendientes de 1 a 1.5%, con escurrimiento lento, la permeabilidad moderadamente lenta y el drenaje es imperfecto a moderado. Son suelos poco profundos (0.60 m) de muy baja fertilidad, con escaso tenor de materia orgánica.

La Serie de suelo Pampín se ubica en relieve normal, posición de loma, con pendientes de 1 a 1,5%. Son suelos profundos y de baja fertilidad, compuestos por un manto arenoso de 120 cm. de espesor.

El uso actual es el de campo natural de pastoreo, forestación y cultivos de hortalizas. El tapiz vegetal está compuesto de *Andropogon lateralis, Paspalum notatum, Sorghastrum setosum, Cynodon sp., Sporobolus sp. Schizachirium sp., y Axonopus sp*.., *Desmodiun incanum* y otros de hábitos húmedos como Ciperáceas y Centella. El principal uso de los pastizales es la ganadería extensiva y la para forestación de bosques cultivados con pino y eucalipto. Cuando se mejoran las condiciones de drenaje y fertilidad se los utiliza para agricultura. Se ubica en la Clase IVw y el índice de Productividad es de 16 (Escobar, et al., 1996). Con la información del mapa de suelo de la provincia de Corrientes e imágenes satelitales, en el área de estudio se identificó las Series de Suelos Chavarria y Pampín.

En cada sitio y a partir de un punto se estableció una transecta para el muestreo de la vegetación. La línea gruesa y central indica la senda a partir de la cual

se muestrea ambos lados de la transecta de 50 m y con el auxilio de un marco de 1.0 x 1.0 m se tomaron 5 muestras de la parte aérea de las especies forrajeras que conforman el tapiz, cortando con una tijera a una altura de 2 cm sobre el suelo. (Fig. 4.2)

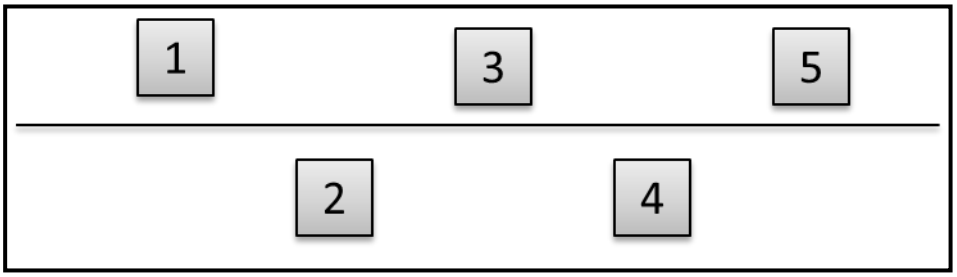


Fig. 4.2. Diseño de la transecta utilizada para el muestreo de la vegetación. La gráfica corresponde a una distancia de 50 m.

Se eliminó el material muerto encontrado. Las muestras fueron tomadas en el período estival y el estado fenológico fue al comienzo de la floración. Las muestras se llevaron al laboratorio y se determinó el peso de la misma y la composición florística.

4.2 Reactivos

Todos los reactivos utilizados fueron de grado analítico. Para las diluciones se utilizó agua desionizada (18 MΩ cm^{-1}) obtenidas utilizando un sistema Mili-Q modelo TM Plus. Las diluciones se prepararon a partir de una solución multielemental estándar IV TraceCERT, otra solución multielemental V TraceCERT y otra solución multielemental II de TraceCERT.

4.3 Determinación Multielemental

4.3.1 Digestión asistida por microondas

En el caso de elementos volátiles a baja temperatura (por ej. As, Hg, Se y sus compuestos) o compuestos refractarios (tierras raras, carburos y otros), el empleo de métodos de digestión asistidos por calentamiento con radiación electromagnéticas de

microondas usando frascos cerrados de teflón es un prerrequisito para evitar pérdidas por volatilización, además de evitar problemas normalmente observados en los procedimientos convencionales de digestión, tales como la posibilidad de descomposición incompleta, etapas largas, exhaustivas y con riesgos de contaminación. La utilización del método de digestión usando radiación de microondas en frasco es especialmente interesante considerando la importancia de estudios con relación a especies potencialmente tóxicas.

Los reactivos normalmente empleados en los procedimientos de digestión de muestras usando horno de microondas son: agua desionizada, ácido nítrico (HNO_3) concentrado, peróxido de hidrógeno (H_2O_2) concentrado, ácido fluorhídrico (HF) concentrado, ácido bórico (H_3BO_3) concentrado, ácido clorhídrico (HCl) concentrado. En el caso del análisis de concentraciones traza, se recomienda la utilización de reactivos de alta pureza o destilados por debajo de su temperatura de ebullición. Aproximadamente 500 mg de muestra limpia se insertó dentro de un recipiente para microondas. Dos mililitros de H_2O_2 al 30% (v/v) y 6 mL de HNO_3 al 65% (m/v) se añadieron en el recipiente. Un programa de temperatura fue aplicado para realizar la digestión (etapa 1:25 a 200 ˚C durante 15 minutos, etapa 2: 200 ˚C durante 15 minutos, etapa 3: 200 a 110 ˚C durante 15 minutos). Finalmente, la muestra digerida se diluyó hasta un volumen final de 10 mL con agua ultrapura. En la Figura 4.3 se muestra un esquema ilustrativo de los pasos seguidos.

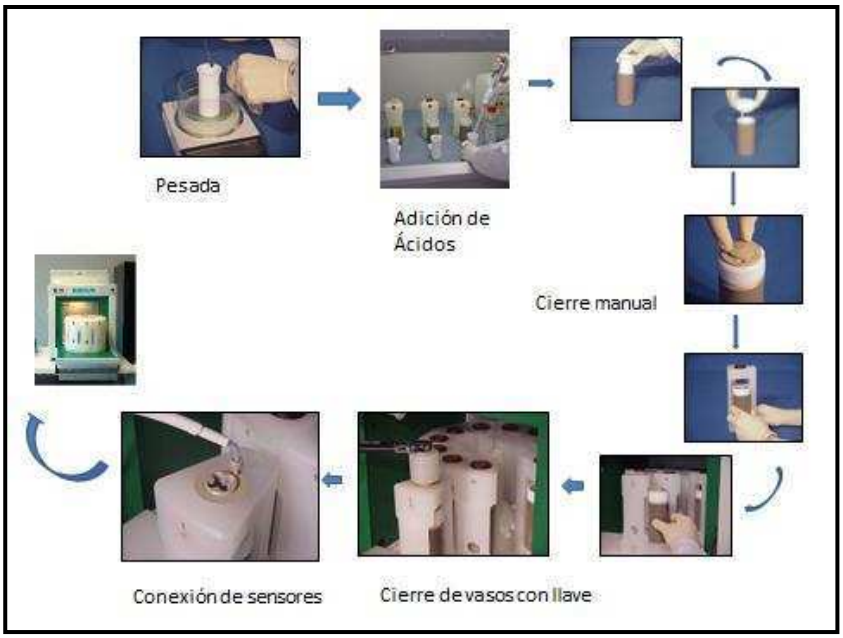


Fig. 4.3 Metodología del horno de microondas

4.3.1.1 Ventajas de la digestión con microondas

- Rápida (aproximadamente 30 minutos).
- Control de la temperatura.
- Bajo riesgo de contaminación, se obtienen mejores blancos.
- No hay pérdida de elementos volátiles durante la digestión (Hg, As, Sb, Se, Sn, B y Cr).
- Procesado de varias muestras al mismo tiempo.
- Reduce el gasto de reactivos ácidos (5-10 mL).
- El operador no está expuesto a vapores tóxicos o peligrosos.
- Se procesa poca cantidad de muestra (0,1-2 g).

4.3.2 Determinación mediante Espectrometría de Masas con Plasma Acoplado Inductivamente (ICP-MS)

Para el análisis multielemental se utilizó un equipo de espectrometría de masas con plasma acoplado inductivamente (ICP-MS) Agilent 7700 Series. Los detalles sobre el funcionamiento del equipo se establecen en la Tabla 4.2.

Tabla 4.2. Condiciones de operación del equipo ICP-MS Agilent 7700 Series

Condiciones de operación	
Nebulizador	Quartz concentric (Micromist) 400 mL/min
Spray chamber	Scott-type double-pass water cooled
Geometría de la celda	Octopolo
Cono de sampling	Nickel, 1.0 mm orific
Skimmer	Nickel, 0.75 mm orific
RF Power	1400–1500 W
Reflected Power	<10 W
Modo estándar	
Flujo de gas	15 L/min
Flujo de gas nebulizador	0.95 – 1.0 L/min
Flujo de gas auxiliar	0.99 L/min
Expansion stage	2.0 mbar

Los isótopos analizados fueron: Al^{+27}, B^{+10}, Cd^{+111}, Co^{+59}, Cr^{+53}, Cu^{+63}, Li^{+7}, Mo^{+95}, Ni^{+62}, Rb^{+85}, Sb^{+121}, Se^{+80}, Sn^{+118}, Sr^{+88}, Ti^{+49}, Tl^{+205}, V^{+51} y Zn^{+66}.

4.3.3 Control de calidad de los resultados

Con el objetivo de controlar la calidad de los resultados, en este trabajo se realizaron los siguientes métodos para la validación de los resultados:

1. Calibración instrumental
2. Determinación de la precisión del método
3. Adición de estándar interno
4. Adición de estándar

4.3.3.1 Calibración instrumental

Al solo efecto de validar el funcionamiento del instrumento se realizaron curvas de calibración para cada elemento a determinar con 5 concentraciones distintas en solución de HNO_3 2% (m/v). En general se intenta que la matriz de los estándares sea lo más similar posible a la matriz de la muestra. Cada blanco y patrón utilizado se leyó mediante la nebulización directa de los mismos. Las diluciones se realizaron a partir de una solución multielemental estándar IV para ICP TRACECERT, una solución multielemental III para ICP TRACECERT, una solución mono-elemental de Sb TRACECERT.

4.3.3.2 Precisión del método

La precisión de un método analítico representa el parámetro que refleja el grado de concordancia que existe entre un conjunto de valores obtenidos al realizar una serie de medidas repetitivas o independientes una de otra bajo condiciones específicas. Teniendo en cuenta las condiciones en que se realizan las medidas, se puede distinguir la reproducibilidad y repetibilidad del método. Con respecto a la repetibilidad, los resultados se obtienen de ensayos mutuamente independientes mediante el mismo método aplicado a la muestra a analizar, en el mismo laboratorio, con el mismo equipamiento y por el mismo operador en un intervalo corto de tiempo. Es una medida de la varianza interna y refleja la precisión máxima que se puede obtener con un dado método analítico. En cuanto a la reproducibilidad, requiere que se obtengan resultados de ensayos independientes mediante el mismo método aplicado a la muestra a analizar en diferentes condiciones como diferentes laboratorios, o diferentes equipos o diferentes operadores. Los estimadores de la

precisión que se han utilizado en este trabajo son la desviación estándar (s), y la desviación estándar relativa (RSD, expresada en %).

4.3.3.3 Adición de estándar interno

En esta técnica se añade un volumen fijo de un elemento elegido como patrón interno tanto en muestras, blancos y patrones. En este trabajo, el elemento elegido fue el In, debido a que este elemento se ioniza casi a un 100% y el mismo es raramente encontrado en alimentos. A continuación, se determinan las respuestas del analito y del estándar interno, y se calcula el cociente de las dos respuestas. Es esperable que si varía algún parámetro que afecte a las respuestas medidas, dichas respuestas (del analito y estándar interno) se verán afectadas en igual proporción.

4.3.3.4 Adición de estándar

Este método consiste en agregar una concentración conocida de estándar de cada elemento a una muestra problema de composición ya determinada. A continuación, se repite el proceso de medida químico y la señal obtenida se deberá a la cantidad de analito originalmente presente en la muestra sumada a la cantidad agregada. Con los resultados obtenidos se calcula el porcentaje de recuperación. El propósito de este método es corregir la presencia de posibles efectos de interferencias debidas a la matriz de la muestra.

4.4 Análisis estadístico de los resultados

Los resultados obtenidos a través del análisis multielemental se presentan mediante el cálculo de la media y desviación estándar en forma de tablas. A continuación, se realizaron gráficos estadísticos tipo gráficos de caja para explicar la dispersión de los mismos. Finalmente, los resultados obtenidos fueron utilizados para

el análisis quimiométrico y modelado según las técnicas descriptas en el capítulo 3. Para ello se utilizó el software R 3.3.1 (Team, R. C. 2016) y los paquetes caret (Kuhn, M. 2016) y mlr (Bernd Bischl, M. L., Jakob Richter, Jakob Bossek, Leonard Judt, Tobias Kuehn, Erich Studerus, Lars Kotthoff 2017).

4.5 Referencias Bibliográficas

mlr: Machine Learning in R [2017.

caret: Classification and Regression Training. R package version 6.0-70 [2016.

R: A Language and Environment for Statistical Computing [2016. Version 3.3.1. Vienna, Austria.

Sección III: RESULTADOS Y DISCUSIÓN

5) Capítulo V

5.1 Optimización y Validación de Resultados

Se confeccionaron rectas de calibración para cada elemento a partir de patrones por triplicado y empleando 5 niveles de concentración. Las diluciones se prepararon a partir de una solución multielemental estándar IV TraceCERT, otra solución multielemental V TraceCERT y otra solución multielemental II de TraceCERT. Los coeficientes de regresión obtenidos para cada recta de calibración tuvieron un coeficiente de determinación (R^2) comprendido entre 0,9980 y 0,9991. Estas rectas de calibración fueron utilizadas para la cuantificación de estos elementos en las muestras.

5.2 Precisión del Método

Por otro lado, se realizaron lecturas replicadas de una solución estándar conteniendo todos los elementos, en un mismo día de trabajo y en distintos días de trabajo, con nuevas condiciones de trabajo de equipo. Estas mediciones se realizaron con el objeto de evaluar la precisión del método, teniendo en cuenta la repetibilidad y la reproducibilidad del mismo. Valores inferiores a 4,5% de RSD se obtuvieron para las mediciones realizadas en un mismo día (repetibilidad) y valores inferiores a 2,5% se alcanzaron para mediciones replicadas en distintos días (reproducibilidad), por lo que se puede afirmar que la precisión de la metodología propuesta es adecuada.

5.3 Adición de estándar interno

Previo a la digestión se agregó a cada muestra ya pesada en las bombas de teflón, 1,0 mL de solución de In (10 mg/L) preparada a partir de una solución de In en ácido nítrico (1000 mg/L), como estándar interno. La concentración de este elemento se determinó luego de manera simultánea al resto de analitos por ICP-MS, y se calculó

el porcentaje de recuperación del mismo. Este procedimiento de agregado de estándar interno permitió evaluar la calidad de los pre-tratamientos aplicados a las muestras y posibles pérdidas que pudieran ocurrir durante las etapas de digestión. El elemento In se seleccionó debido a que en ensayos cualitativos previos practicados en tres muestras problemas seleccionadas al azar, dicho elemento se encontraba en niveles no detectables. Como resultado de estas experiencias se obtuvo un valor promedio de 96,7% ± 3,7% de recuperación en 10 muestras seleccionadas al azar, por lo que se puede afirmar que el pre-tratamiento aplicado a las muestras resulta adecuado para los niveles de exactitud requeridos en este trabajo.

5.4 Adición de estándar

En la tabla 5.1, se observan los resultados obtenidos para la evaluación del grado de recuperación de cada uno de los elementos analizados, empleando el método de adición estándar, el cual es considerado como un método adecuado para evaluar la exactitud del método aplicado sobre muestras de matriz compleja. También se presentan los límites de detección (LD) que es la mínima concentración detectable de manera confiable por la técnica analítica.

Tabla 5.1 Prueba de adición estándar

Elementos	LD (µg/g)	Recuperación (%)
Al[a]	0,092	98,2
B	0,500	103,5
Ba	0,090	96,2
Co	0,002	101,0
Cr	0,010	103,7
Cu	0,022	105,2
Li	0,005	99,1
Mo	0,018	98,9
Ni	0,058	100,1
Rb	0,019	108
Sb	0,001	109,3
Se	0,010	95,0
Sn	0,007	101,5
Sr[a]	0,013	96,9
Ti	0,052	103,1
Tl	0,012	112,5
V	0,020	102,3
Zn	0,100	97,8

[a] Elementos repicados a 100 ng/g. Los otros elementos fueron replicados a 10 ng/g.

Como se puede observar los porcentajes de recuperación para los elementos se encontraron entre 95 % (Se) y 109% (Sb), lo que indica que el método propuesto es adecuado para cuantificar los analitos en la matriz de las muestras problema, desde el punto de vista del porcentaje de recuperación de los mismos.

5.5 Resultados experimentales

5.5.1 Desmodium incanum

Los resultados obtenidos para los 18 elementos (Al, B, Cd, Co, Cr, Cu, Li, Mo, Ni, Rb, Sb, Se, Sn, Sr, Ti, Tl, V y Zn) analizados en todas las muestras (*n* = 42), agrupadas según series de suelo (Chavarría y Pampín) estudiadas se presentan en la tabla 5.2.

Tabla 5.2 Concentraciones de los elementos minerales en muestras de partes aéreas de *Desmodium incanum* (media ± desvío estándar)

Elemento	Chavarría (µg/g)	Pampín (µg/g)
Al	0,23 ± 0,12	0,35 ± 0,12
B	0,4 ± 0,1	2,3 ± 1,4
Cd	0,05 ± 0,01	0,05 ± 0,01
Co	0,02 ± 0,01	0,04 ± 0,01
Cr	0,05 ± 0,01	0,06 ± 0,01
Cu	3,7 ± 0,4	3,2 ± 1,0
Li	3,2 ± 1,1	2,6 ± 0,8
Mo	0,22 ± 0,01	0,25 ± 0,01
Ni	0,04 ± 0,03	0,04 ± 0,02
Rb	0,74 ± 0,33	2,3 ± 1,1
Sb	0,02 ± 0,01	0,02 ± 0,01
Se	0,08 ± 0,01	0,06 ± 0,01
Sn	0,06 ± 0,01	0,06 ± 0,02
Sr	7,3 ± 2,3	9,3 ± 2,8
Ti	2,4 ± 0,5	2,3 ± 0,5
Tl	0,04 ± 0,01	0,03 ± 0,01
V	0,04 ± 0,01	0,16 ± 0,1
Zn	9,4 ± 3,1	13,8 ± 2,7

Adicionalmente, los resultados se representan a continuación utilizando gráficos de cajas y bigotes o boxplot. Un gráfico de Cajas y bigotes permite visualizar la distribución de valores de una determinada variable cuantitativa en estudio. En este caso se aplicó esta técnica de estadística descriptiva para una mejor visualización y comprensión de las variables. Se agruparon las variables según tengan un rango similar de valores. Se visualizan los valores de Al, Mo y V en la Fig. 5.1. Se observó que la distribución de valores en las muestras de Pampín es superior a las de Chavarría, pero con respecto al V; la dispersión es mayor en las muestras de Pampín. Las medias también de Al y Mo son similares en Pampín, lo que no sucede en las muestras de Chavarría. Los resultados de Mo están en rangos similares a los resultados presentados por Gupta para alfalfa; y son un poco menores (0,64 mg/kg) a los valores presentados por Watson para *Perennial ryegrass* (Gupta, U. C. et al. 2008, Watson, C. A. et al. 2012).

Los valores de molibdeno estuvieron entre 0,19 y 0,52 µg/g para las especies de D. incanum recolectadas en Pampín, y entre 0,16 y 0,48 µg/g para las recolectadas en Chavarría. Los valores de este trabajo están por encima de los valores críticos (0,1 µg/g) para ganado productor de carne. Según reportes, es poco probable que ocurra una deficiencia de Mo en nuestro país, según las condiciones de pastoreo. Si bien los valores reportados en este trabajo están por debajo del máximo tolerable (5-6 mg/kg materia seca) para ganado vacuno productor de carne, es importante tener en cuenta la toxicidad del Mo para el ganado vacuno. En estos casos, hay que tener en cuenta las concentraciones de Cu ya que éste actúa como protector ante la toxicidad del molibdeno (Bernardis, A., et al. 2016, Mufarrege, D. 1999).

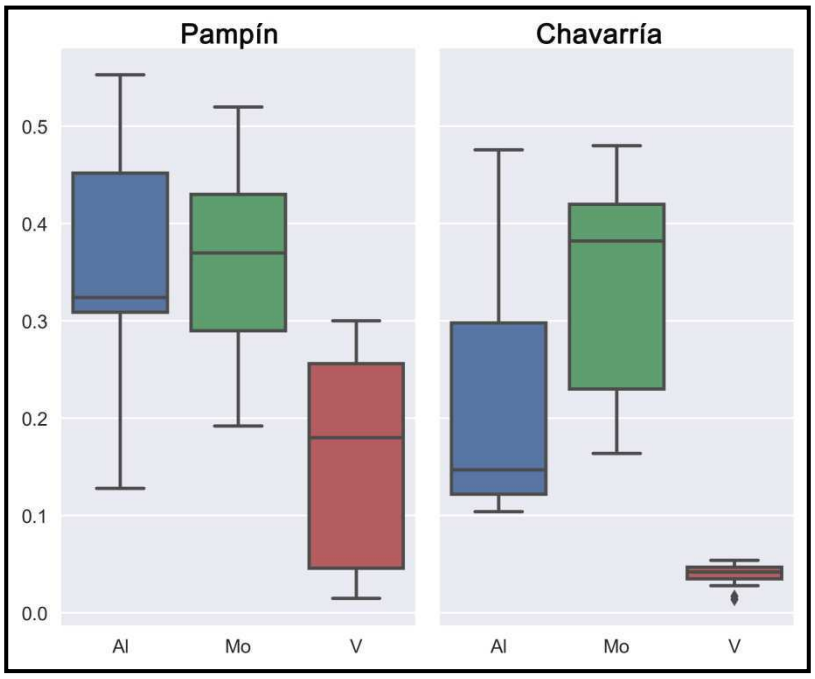


Fig.5.1 Gráficos Cajas y Bigotes de Al, Mo y V en muestras de *Desmodium incanum* en las dos series de suelo

En la Fig. 5.2 se observan los gráficos de cajas y bigotes pertenecientes a B y Rb de *Desmodium incanum* pertenecientes a las series de suelo: Chavarría y Pampín. La dispersión de valores es mucho mayor en las muestras de Pampín, mientras que la

dispersión en las muestras de Chavarría es pequeña. Estos valores de B son inferiores a los presentados por Gupta y Watson para países nórdicos y Reino Unido (Gupta, U. C., et al. 2008, Watson, C. A., et al. 2012).

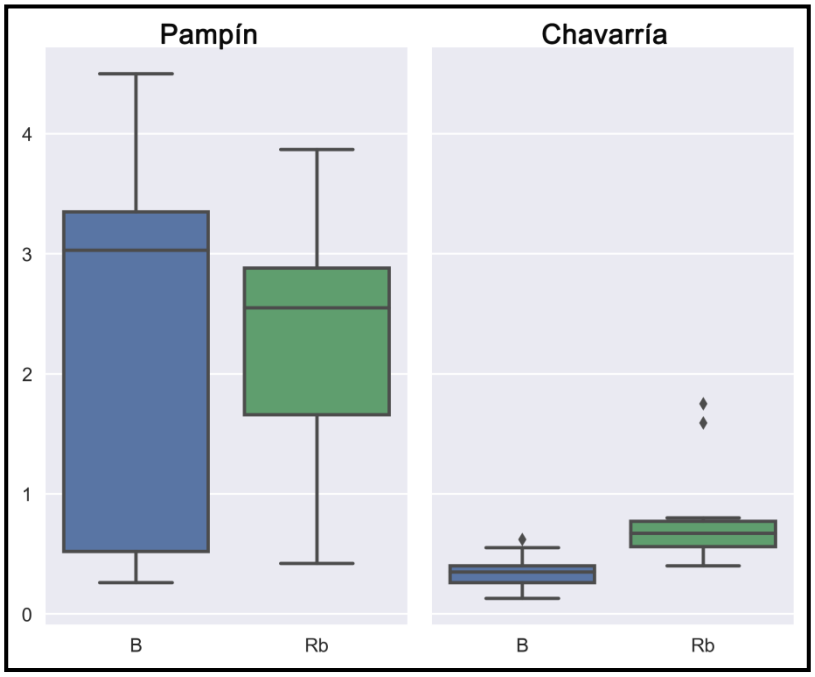


Fig.5.2 Gráficos de Cajas y Bigotes de B y Rb en muestras de *Desmodium incanum* en dos series de suelo

En la Fig. 5.3 se observan la distribución de valores de Cd, Co, Se, Sb, Sn, Cr, Ni y Tl de *Desmodium incanum* en las dos series de suelo. Los valores de Cd prácticamente no tienen dispersión alguna, aun así, se observan valores extremos en ambas series de suelo. Lo mismo sucede con los valores de Co en la serie de suelo Chavarría, pero las muestras de Pampín tienen mayor dispersión. *La concentración de cobalto no cubre el requerimiento (1 µg/g) para ganado productor de carne. Los requerimientos de los bovinos para carne quedan satisfechos con 0,10 ppm de Co en materia seca de la dieta. Esta deficiencia es común en regiones tropicales, como así también, se han detectado en pasturas de la región pampeana. En caso de sospecharse una deficiencia de Co, se deberían hacerse ensayos en animales suplementados, en potreros diferentes. La*

deficiencia en vacunos puede ser detectada por análisis de Co ó de Vitamina B12 en suero ó hígado, siendo recomendable la determinación de B12 en hígado. La suplementación con Co puede hacerse mediante suplementos minerales con Sulfato ó Carbonato de Cobalto (Bernardis, A., et al. 2016, Mufarrege, D. 1999).

Con respecto al Se, los valores medios son más altos en la serie Chavarría, aunque están en rangos prácticamente similares. *Los valores de selenio en este trabajo están por debajo de los valores de los requerimientos para ganado vacuno productor de carne (0,1 µg/g). El selenio es importante en diversas funciones corporales, como el crecimiento, reproducción, prevención de enfermedades y la integridad de tejidos. En caso de deficiencia, la suplementación se puede hacer mediante cantidades pequeñas, de unos 1,5 mg de selenito de sodio en mezclas minerales en la ración de vacunos (Bernardis, A., et al. 2016, Mufarrege, D. 1999).* Los valores de Sb, Cr y Ni son bastante similares entre sí en cuanto a valores medios y dispersión en ambas series de suelo, no sucede así con Sn que en Chavarría tienen mayor dispersión de valores*. Los valores de cromo en este trabajo están por debajo del requerimiento para ganado productor de carne (1 µg/g). En caso de deficiencia la adición de Cr en forma de picolinato de cromo (0,05 mg/kg) a la dieta de terneros en crecimiento, aumentó la tasa de desaparición de glucosa, ya que el Cr está involucrado en potenciar la acción de la insulina (Bernardis, A., et al. 2016, Mufarrege, D. 1999).*

Con respecto a los valores de Tl, si se observan diferencias entre los valores de ambas series, pero existe una mayor dispersión en los valores de Chavarría, como así también unos valores medios y mediana superiores.

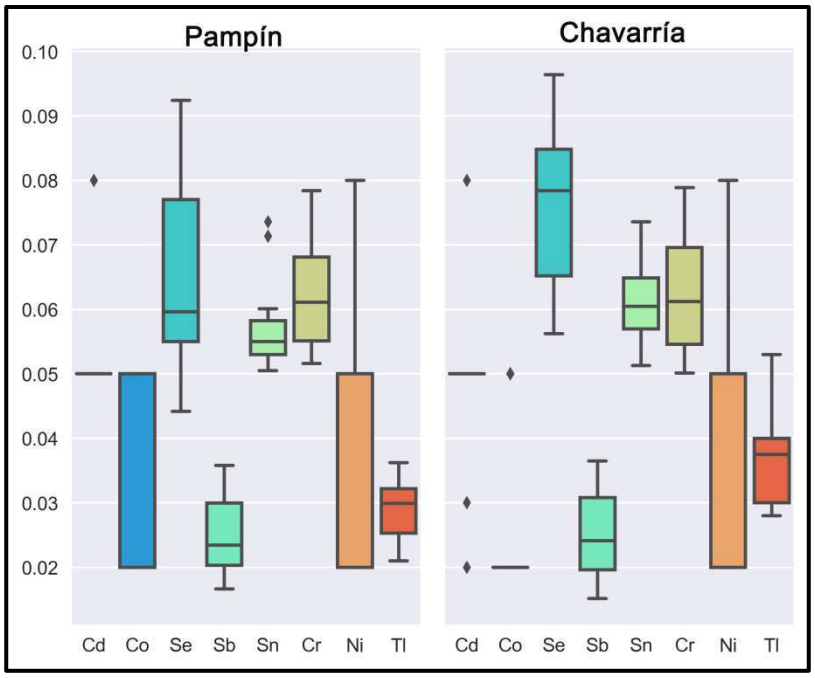


Fig.5.3 Gráficos de Cajas y Bigotes de Cd, Co, Se, Sb, Sn, Cr, Ni y Tl en muestras de *Desmodium incanum* en dos series de suelo

En la Fig. 5.4 se observan los valores correspondientes a Li, Cu y Ti en las muestras de *Desmodium incanum* provenientes de Chavarría y Pampín. Los valores de Li presentan mayor dispersión en la serie de Chavarría, con valores medios superiores en Chavarría. Con respecto a Cu, presentan un valor medio superior en Chavarría, aunque existe mayor dispersión es en la serie Pampín. Estos valores de Cu están en los mismos rangos de valores presentados por Moscuzza, y levemente inferiores a los presentados por Gupta y Watson (Gupta, U. C., et al. 2008, Moscuzza, C. H. et al. 2012, Watson, C. A., et al. 2012). *Los valores de cobre no cubren el requerimiento para ganado productor de carne (10 µg/g). La falta de cobre en vacunos se caracteriza por diferentes trastornos, entre ellos, lento crecimiento, reducción de la fertilidad, quebraduras espontaneas en animales jóvenes, diarrea y anemia, y la producción de anticuerpos. Los requerimientos de cobre siempre dependen de la concentración de*

molibdeno, sulfatos inorgánicos y hierro en el alimento (Bernardis, A., et al. 2016, Mufarrege, D. 2003).

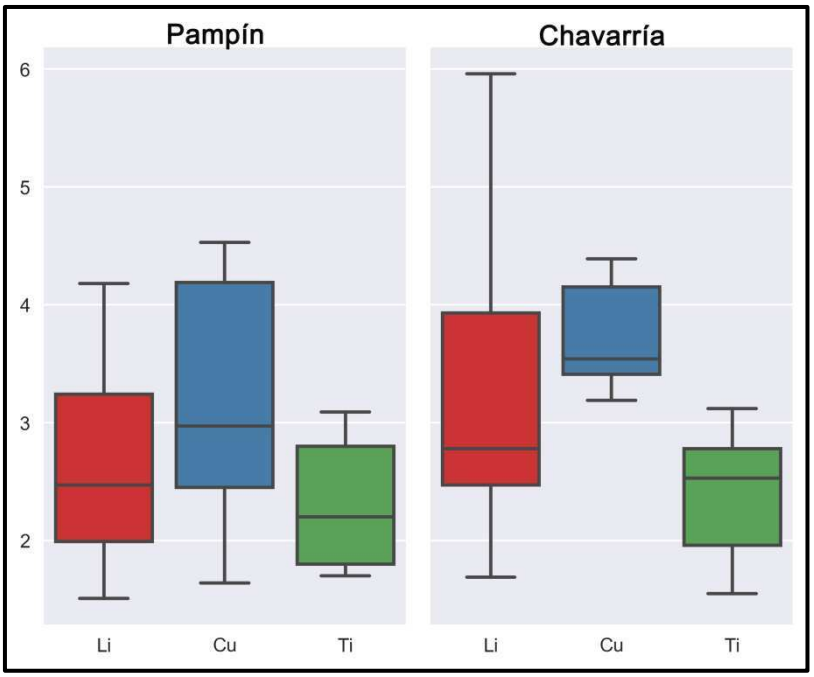


Fig. 5.4 Gráficos de Cajas y Bigotes de Li, Cu y Ti en muestras de *Desmodium incanum* en dos series de suelo

En la Fig. 5.5 se observan los valores de Sr y Zn para muestras de *Desmodium incanum* en las series de suelo; Chavarría y Pampín. Los valores de Zn presentan una distribución más grande como así también una media y mediana mayor en las muestras de Pampín. La media y mediana de Zn en las muestras de Pampín están bastante próximas y sucede lo mismo con las muestras de Sr de Chavarría. Los valores de Zn están por debajo de los valores que se consideran con deficiencia, y por debajo de los valores reportados por Moscuzza, Watson y Gupta (Gupta, U. C., et al. 2008, Moscuzza, C. H., et al. 2012, Watson, C. A., et al. 2012*). Los valores de zinc en este trabajo están por debajo del requerimiento en la mayoría de los casos (20 µg/g) para ganado productor de carne. Otros estudios de tejidos corporales de animales con y sin*

implementación de elementos minerales, demostraron que el Zn no tiene en el organismo un tejido de reserva y fácil acceso, y si se produce escasez en el pastoreo, la deficiencia de Zn comienza a suceder. Para corregir esta deficiencia, lo que se realiza en incorporar 0,5% de Zn en mezclas minerales que se utilizan como suplemento, considerándose esto como suficiente para corregir cualquier probable deficiencia (Bernardis, A., et al. 2016, Mufarrege, D. 2000).

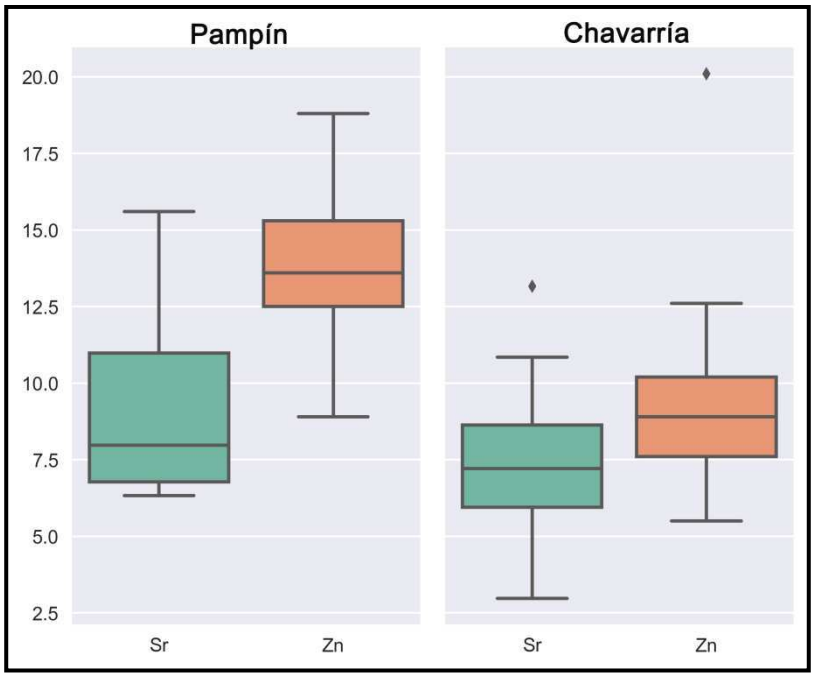


Fig. 5.5 Gráficos de Cajas y Bigotes de Sr y Zn en muestras de *Desmodium incanum* en dos series de suelo

5.5.2 *Schizachyrium microstachyum*

Los resultados obtenidos para los 18 elementos (Al, B, Cd, Co, Cr, Cu, Li, Mo, Ni, Rb, Sb, Se, Sn, Sr, Ti, Tl, V y Zn) analizados en todas las muestras (*n* = 48), agrupadas según series de suelo (Chavarría y Pampín) estudiadas se presentan en la tabla 5.3.

Tabla 5.3 Concentraciones de los elementos minerales en muestras de partes aéreas de *Schizachyrium microstachyum* (media ± desvío estándar)

Elemento	Chavarría (µg/g)	Pampín (µg/g)
Al	0,09 ± 0,01	0,10 ± 0,02
B	3,12 ± 0,6	2,84 ± 0,7
Cd	0,05 ± 0,01	0,05 ± 0,02
Co	0,04 ± 0,01	0,03 ± 0,01
Cr	0,07 ± 0,01	0,05 ± 0,02
Cu	2,4 ± 0,8	3,7 ± 0,6
Li	2,3 ± 0,6	3,2 ± 1,9
Mo	0,11 ± 0,03	0,14 ± 0,03
Ni	0,04 ± 0,02	0,03 ± 0,02
Rb	1,4 ± 0,2	1,2 ± 0,2
Sb	0,03 ± 0,01	0,03 ± 0,01
Se	0,07 ± 0,01	0,08 ± 0,02
Sn	0,11 ± 0,01	0,12 ± 0,01
Sr	14,1 ± 2,0	12,2 ± 2,1
Ti	2,3 ± 0,5	2,4 ± 0,5
Tl	0,02 ± 0,01	0,03 ± 0,01
V	0,22 ± 0,05	0,17 ± 0,07
Zn	14,0 ± 5,3	15,5 ± 3,4

A continuación, se presentan los gráficos de Cajas y Bigotes para una mejor comprensión de la distribución de valores de *Schizachyrium microstachyum* en las dos series de suelo estudiadas. En la Fig. 5.6 se presentan la distribución de valores de Al, Mo y V. Los valores de Al presentan mayor dispersión en las muestras de Pampín, aunque la mediana y la media coinciden en las muestras de Chavarría. Con respecto a Mo, los valores medios y mediana son superiores en Pampín. También es importante destacar que dentro de las muestras de Chavarría presentan un par de valores extremos superiores al cuartil más alto. Estos valores de Mo están en concordancia con los valores de Gupta para alfalfa, pero son inferiores a los valores reportados por Watson para *Perennial ryegrass* (Gupta, U. C., et al. 2008, Watson, C. A., et al. 2012).

Los valores medios y la mediana para V son superiores en Chavarría; la mediana y la media son más similares entre sí, a diferencia de Pampín que son más diferentes. Existe una mayor dispersión en los valores de Pampín, en comparación con las de Chavarría. *Los valores de este trabajo están por encima de los valores críticos (0,1 µg/g) para ganado productor de carne. En general los requerimientos de molibdeno son bajos para los animales. Según Mufarrege, es poco probable que ocurran deficiencias en las condiciones de pastoreo en la Argentina (Bernardis, A., et al. 2016, Mufarrege, D. 1999).*

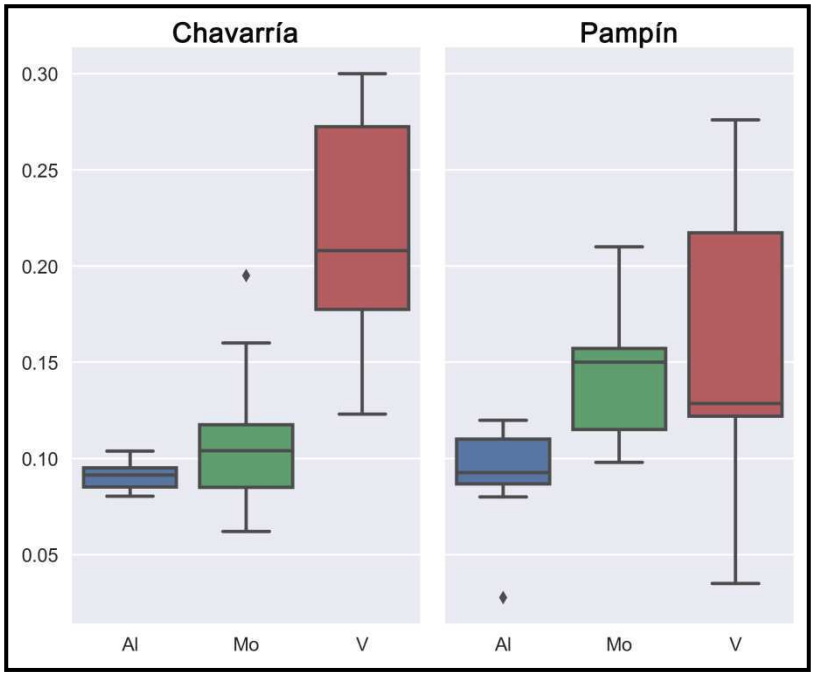


Fig. 5.6 Gráficos de Cajas y Bigotes de Al, Mo y V de *Schizachyrium microstachyum* en dos series de suelo

En la Fig. 5.7 se observan la distribución de valores de B y Rb en *Schizachyrium microstachyum* en dos series de suelo. Se observa que los valores de B tienen mayor dispersión en las muestras de Chavarría que en las de Pampín. En el caso particular de las muestras de Pampín se observa que existen dos valores extremos por debajo y por encima del primer y cuarto cuartil, respectivamente. Estos valores de B están en

niveles inferiores con respecto a los valores reportados tanto por Watson (19 µg/g) y Gupta (14-25 µg/g) (Gupta, U. C., et al. 2008, Watson, C. A., et al. 2012).

Con respecto al Rb la dispersión es un poco mayor en las muestras de Chavarría, aunque los valores medios y la mediana están más próximos.

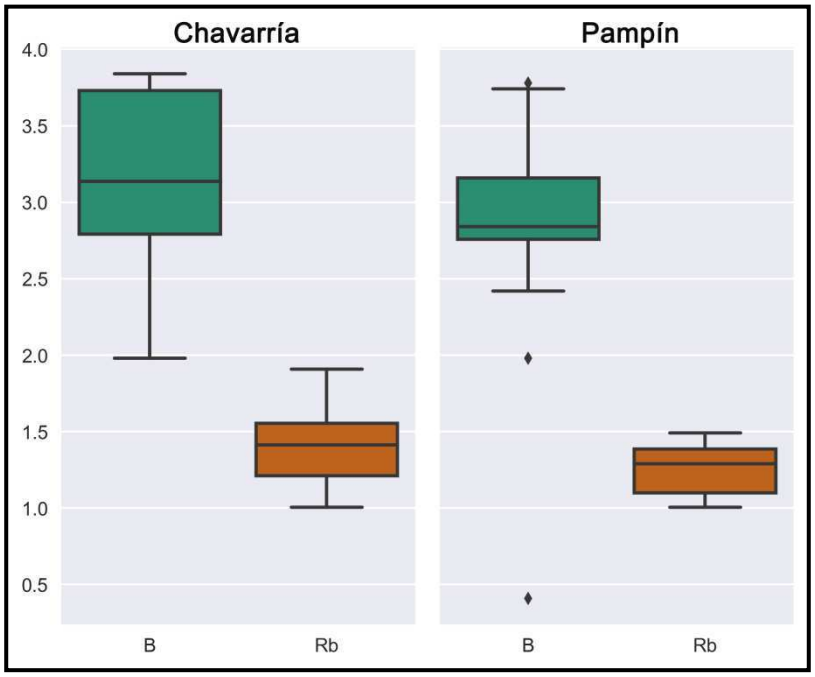


Fig. 5.7 Gráficos de Cajas y Bigotes de B y Rb de *Schizachyrium microstachyum* en dos series de suelo

En la Fig. 5.8 se observa los gráficos de cajas correspondientes a Cd, Co, Se, Sb, Sn, Cr y Ni. Los valores de Cd prácticamente no presentan variación alguna, y se observa sólo una línea en la Fig. 5.8. Sucede lo mismo con los valores de Co y Ni para Pampín, no así con los valores para Chavarría que presentan mayor dispersión. *La concentración de cobalto no cubre el requerimiento (1 µg/g) para ganado productor de carne. En caso de sospechar deficiencias, la suplementación con Co puede hacerse mediante suplementos minerales con Sulfato ó Carbonato de Cobalto (Bernardis, A., et al. 2016, Mufarrege, D. 1999).* Los valores correspondientes a Se presentan mayor

dispersión para las muestras de Pampín, como así también mayor valor medio y mediana. *Los valores de selenio para este trabajo están por debajo de los valores de* los *requerimientos para ganado vacuno productor de carne en materia seca (0,1 µg/g). Una forma de suplementar Se en la dieta se puede hacer agregándolo en forma de un núcleo mineral (mezclas minerales en la ración de los vacunos), ya que la cantidad aproximada a suministrar por animal es muy pequeña, unos 1,5 mg de selenito de sodio (Bernardis, A., et al. 2016, Mufarrege, D. 1999).* Las concentraciones de Sb son bastante similares en ambos casos, al igual que el Cr. *Los valores de cromo en este trabajo están por debajo del requerimiento para ganado productor de carne (1 µg/g). Los resultados obtenidos indican que puede ser necesaria la suplementación de Cr para los bovinos productores de carne. En general, el máximo tolerable es de 1000 mg/kg, y la adición ocurre en forma de sales trivalentes, ya que las sales hexavalentes son más tóxicas (Bernardis, A., et al. 2016, Mufarrege, D. 1999).*

Los valores de Sn presentan mayor dispersión en las muestras de Pampín, como así también el valor medio y la mediana. Las concentraciones de Tl presentan mayor dispersión en las muestras de Chavarría.

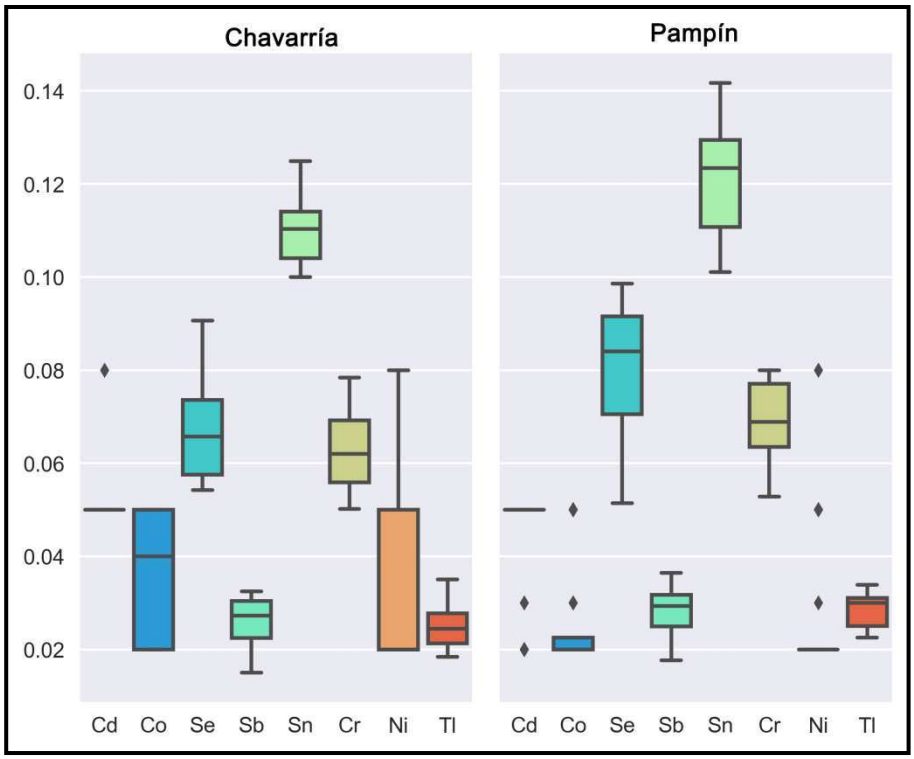


Fig. 5.8 Gráficos de Cajas y Bigotes de Cd, Co, Se, Sb, Sn, Cr, Ni y Tl de *Schizachyrium microstachyum* en las dos series de suelo

En la Fig. 5.9 se presentan los resultados correspondientes a Li, Cu y Ti. En los valores de Li se observa que tienen mayor dispersión los valores de Pampín, como así también se observan valores extremos o outliers, las muestras de Chavarría en cambio presentan menor dispersión y una media y una mediana casi iguales. Con respecto a los valores de Cu, presentan mayor dispersión en Chavarría y valores extremos, en cambio las muestras de Pampín, presentan menor dispersión, mayores valores de media y mediana, y outliers en la zona por debajo del primer cuartil. *Los valores de cobre no cubren el requerimiento para ganado productor de carne (10 µg/g). En general, en la zona del NEA se reportaron deficiencias de cobre, y en general, en todo el territorio argentino. El estudio de cobre en pasturas, es útil para detectar deficiencias en el ganado, debiéndose complementar con S, Mo y Fe en las pasturas y analizar Cu en hígado y en sangre (Bernardis, A. et al. 2016, Mufarrege, D. 2003).* Con respecto a

los valores de Ti se observa una dispersión similar en ambos casos, aunque es ligeramente mayor en Chavarría. En ambos casos, no existen valores extremos para los valores de Ti ni para Chavarría ni para Pampín.

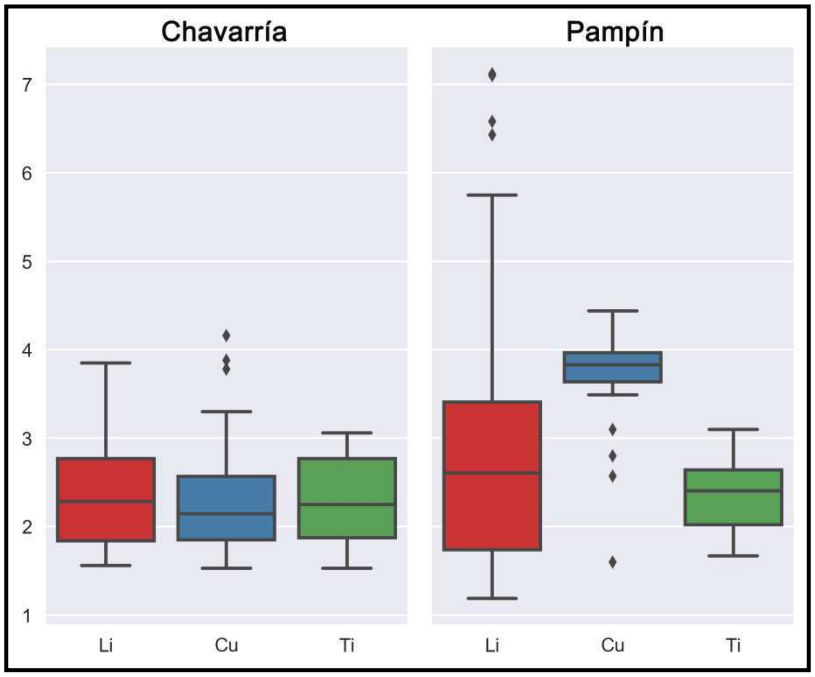


Fig.5.9 Gráficos de Cajas y Bigotes de Li, Cu y Ti de *Schizachyrium microstachyum* en las dos series de suelo

En la Fig. 5.10 se observan las distribuciones de valores de Sr y Zn para *Schizachyrium microstachyum* en Chavarría y Pampín. La mayor dispersión de valores se observa en las muestras de Chavarría, para los valores del Zn, aunque el valor medio y la mediana es superior en los valores de Pampín. La dispersión es menor también para el Sr en las muestras de Pampín, y la media y la mediana son superiores en las muestras de Chavarría. No se observan valores extremos ni para en Sr ni el Zn. Estos valores de Zn están en el mismo rango de los valores reportados por Gupta y de Watson, pero levemente inferiores a los reportados por Moscuzza (Gupta, U. C., et al. 2008, Moscuzza, C. H., et al. 2012, Watson, C. A., et al. 2012). *Los valores de Zn de este trabajo están por debajo del requerimiento en la mayoría de los casos (20 µg/g) para*

ganado productor de carne. La corrección de la deficiencia de Zn en el ganado bovino para carne y en los ovinos, puede hacerse mediante el agregado de 1% de óxido de zinc ó de 3% de sulfato de zinc, a las mezclas minerales suministran normalmente al ganado (Bernardis, A., et al. 2016, Mufarrege, D. 2000).

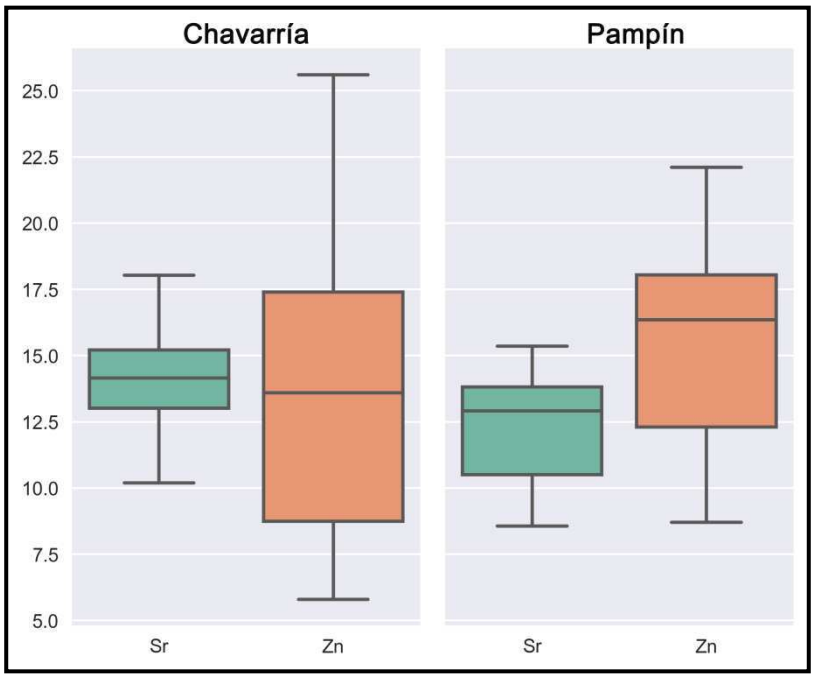


Fig. 5.10 Gráficos de Cajas y Bigotes de Sr y Zn de *Schizachyrium microstachyum* en las dos series de suelo

5.5.3 Andropogon lateralis

Los resultados obtenidos para los 18 elementos analizados (Al, B, Cd, Co, Cr, Cu, Li, Mo, Ni, Rb, Sb, Se, Sn, Sr, Ti, Tl, V y Zn) en todas las muestras (*n* = 43), agrupadas según series de suelo (Chavarría y Pampín) estudiadas se presentan en la tabla 5.4.

Tabla 5.4 Concentraciones de los elementos minerales en muestras de partes aéreas de *Andropogon lateralis* (media ± desvío estándar).

Elemento	Chavarría (µg/g)	Pampín (µg/g)
Al	0,9 ± 0,2	0,9 ± 0,1
B	2,8 ± 0,7	3,1 ± 0,6
Cd	0,05 ± 0,01	0,05 ± 0,01
Co	0,02 ± 0,01	0,04 ± 0,01
Cr	0,07 ± 0,01	0,06 ± 0,01
Cu	3,7 ± 0,5	2,5 ± 1,0
Li	3,6 ± 1,9	2,0 ± 0,5
Mo	0,29 ± 0,07	0,22 ± 0,05
Ni	0,02 ± 0,01	0,04 ± 0,02
Rb	0,62 ± 0,08	0,67 ± 0,12
Sb	0,02 ± 0,01	0,02 ± 0,01
Se	0,07 ± 0,02	0,07 ± 0,01
Sn	0,03 ± 0,01	0,03 ± 0,01
Sr	11,9 ± 2,1	14,1 ± 1,7
Ti	2,3 ± 0,5	2,3 ± 0,4
Tl	0,02 ± 0,01	0,03 ± 0,01
V	0,16 ± 0,06	0,22 ± 0,05
Zn	14,4 ± 3,6	15,6 ± 5,1

En las Fig. 5.11 a Fig. 5.14 se presentan los gráficos de Cajas y Bigotes para *Andropogon lateralis* en dos series de suelo. En la Fig. 5.11 se presenta la distribución de valores de Mo, Al y Rb. Se observa que los valores de Mo presentan baja dispersión en las muestras de Pampín, pero en ambos casos la media y la mediana coinciden. Estos valores de Mo son superiores a los valores reportados por Watson y Gupta (Gupta, U. C., et al. 2008, Watson, C. A., et al. 2012). *Los valores de Mo este trabajo están por encima de los valores críticos (0,1 µg/g) para ganado productor de carne. Mufarrege presenta valores en el mismo orden pero superiores (0,8 mg/kg) en materia seca de pasturas del NEA (Mufarrege, D. 1999). En la provincia de Chaco, sin embargo, se reportan valores de 3 a 42 mg/kg para Melilotus albus (Balbuena O, et al. 2013).*

En cuanto a los valores de Al, se observa mayor dispersión en las muestras de Chavarría, aunque la media y la mediana son iguales. Para las muestras de Pampín en

cambio, la media es mayor que la mediana y se observa un valor extremo por encima del último cuartil. Con respecto a los valores de Rb, se observa mayor dispersión en las muestras de Pampín, aunque la media y la mediana coinciden y son ligeramente mayores que las de Chavarría. Los valores de V presentan mayor dispersión en las muestras de Chavarría, pero la media es mayor en Pampín

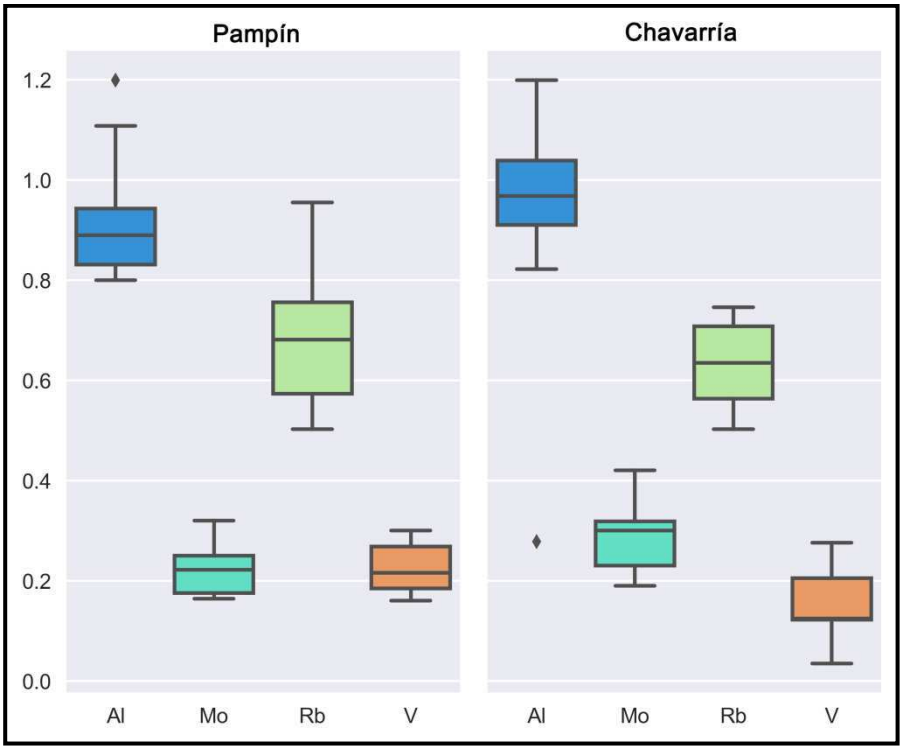


Fig. 5.11 Gráficos de Cajas y Bigotes de Al, Mo, Rb y V en *Andropogon lateralis* de las dos series de suelo

En la Fig. 5.12 se presentan las distribuciones de valores de Cd, Co, Se, Sb, Sn, Cr, Ni y Tl de *Andropogon lateralis* en dos series de suelo. Con respecto a las muestras de Cd, la dispersión observada es mínima, y se representa mediante una línea. Lo mismo sucede con las muestras de Co y Ni para Chavarría, en cambio en Pampín se observa una gran dispersión de valores de Co y Ni. Estos valores de Co están por debajo de los valores reportados por Watson (Watson, C. A., et al. 2012). *La concentración de cobalto no cubre el requerimiento (1 µg/g) para ganado vacuno*

productor de carne. Valores similares fueron obtenidos por Mufarrege (Mufarrege, D. 1999) en tanto que Balbuena (Balbuena O, et al. 2013) obtuvieron valores que oscilan entre 0,07 y 0,24 mg/kg. Para el caso del Se, se observa mayor dispersión en las muestras de Chavarría, y una media y una mediana superior que las muestras de Pampín. Estos valores de Se están por arriba de los valores reportados por Watson (Watson, C. A., et al. 2012). *Los valores de selenio en este trabajo están por debajo de los valores de los requerimientos en materia seca (0,1 µg/g) para ganado vacuno productor de carne (Bernardis, A., et al. 2016). La deficiencia de este mineral afecta la reproducción incluyendo la retención de placenta, y en los terneros produce la enfermedad del "músculo blanco" en los terneros, caracterizada por debilidad, rigidez y deterioro de los músculos en los animales, dificultando que se puedan mantener en pie (Mufarrege, D. 1999).* Las muestras de Chavarría presentan mayor dispersión para los niveles de Sb, aunque se observa un valor extremo por debajo del primer cuartil. Con respecto a los valores de Sn, se observa mayor dispersión en las muestras de Chavarría, y un par de valores extremos para las concentraciones de Pampín. Con respecto al Cr, se observa que la media y la mediana coinciden en las concentraciones de Pampín, y también la dispersión es mayor en Pampín. *Los valores de cromo en este trabajo están por debajo del requerimiento (1 µg/g) para ganado vacuno productor de carne. El papel fisiológico predominante del Cr es integrar el factor de tolerancia a la glucosa que potencia la insulina (Bernardis, A., et al. 2016).*

Con respecto a Tl, la dispersión es mayor en las muestras de Pampín, y la media es mayor a la mediana. En cambio, en las muestras de Chavarría, la mediana es superior a la media, y no se observan valores extremos.

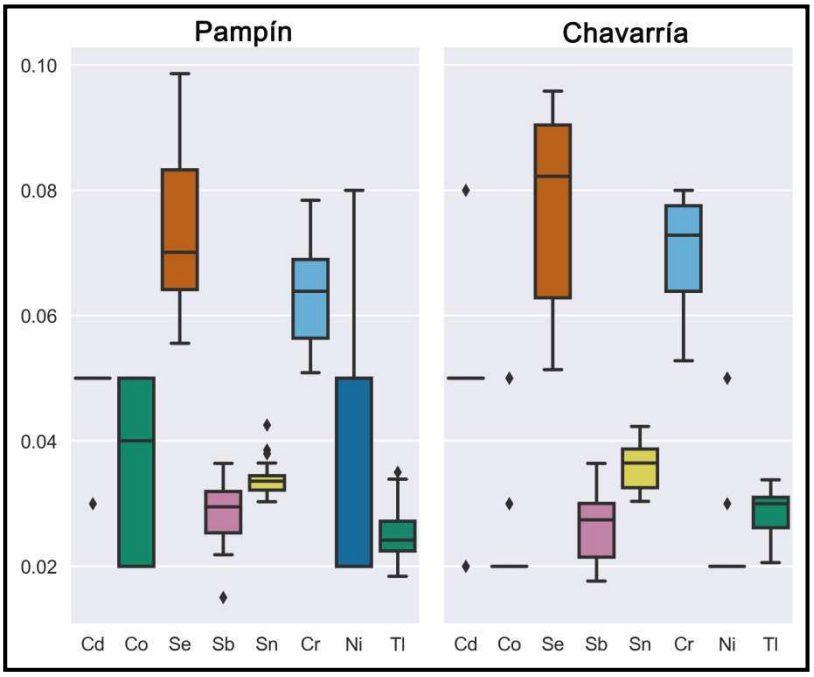


Fig. 5.12 Gráficos de Cajas y Bigotes de Cd, Co, Se, Sb, Sn, Cr, Ni y Tl en *Andropogon lateralis* de dos series de suelo

En la Fig. 5.13 se presenta la distribución de valores de Li, Cu, B y Ti en *Andropogon lateralis* en dos series de suelo. Con respecto a los valores de Li, se observa una dispersión mayor en las muestras de Chavarría, como así también un valor extremo por encima del cuarto cuartil. Los valores correspondientes a Cu presentan una mayor dispersión en Pampín, aunque las muestras de Chavarría presentan valores extremos por encima y por debajo del primer y tercer cuartil. En ninguno de los dos casos, coinciden la media y la mediana. Estos valores de Cu están en los rangos de valores que reporta Moscuzza, pero son menores a los valores reportados con Watson y Gupta (Gupta, U. C., et al. 2008, Moscuzza, C. H., et al. 2012, Watson, C. A., et al. 2012*). Los valores de cobre no cubren el requerimiento (10 µg/g) para ganado vacuno*

productor de carne. La deficiencia de Cu en los forrajes se presenta cuando los suelos tienen deficiencia natural de Cu y por interacciones con otros elementos como Fe, Zn, Cd, Mo y S. El contenido de cobre en las pasturas varía con el tipo de suelo (pH, contenido de materia orgánica), especie de planta, estado de madurez, manejo y clima (Bernardis, A. et al. 2016, Mufarrege, D. 2003).

Con respecto a las muestras de B, se observa mayor dispersión en las muestras de Pampín, aunque las muestras de Chavarría presentan valores extremos por debajo del primer cuartil y el tercer cuartil. La distribución de valores de Pampín son bastante similares y en ambos casos, la media es superior a la mediana. Los valores de B son menores a los reportados por Gupta y Watson (Gupta, U. C., et al. 2008, Watson, C. A., et al. 2012)

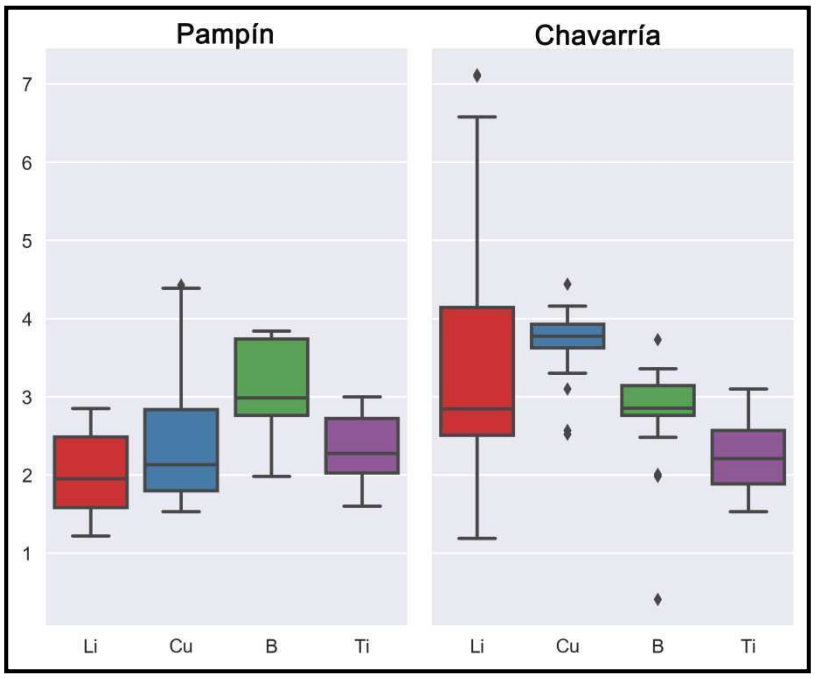


Fig. 5.13 Gráficos de Cajas y Bigotes de Li, Cu, B y Ti en *Andropogon lateralis* de dos series de suelo

En la Fig. 5.14 se presentan los valores de Sr y Zn, se observa mayor dispersión en los valores de Zn de Pampín, aunque la media y la mediana coinciden. En las

muestras de Chavarría, la mediana es superior a la media y la dispersión de valores es menor. Con respecto a los valores de Sr, se observa mayor dispersión en las muestras de Chavarría, aunque los valores de Pampín presentan un valor extremo. Estos valores de Zn son levemente inferiores a los reportados por Moscuzza, Watson y Gupta que reportan valores medios por sobre los 20 mg/kg (Gupta, U. C., et al. 2008, Moscuzza, C. H., et al. 2012, Watson, C. A., et al. 2012). *Los valores de este trabajo están por debajo del requerimiento en la mayoría de los casos (20 µg/g) para ganado vacuno productor de carne. Esto confirma lo reportado por Mufarrege y Bernardis, ya que la región del NEA la concentración de Zn es inferior a 20 mg/kg de materia seca de los pastos naturales (Bernardis, A., et al. 2016, Mufarrege, D. 2000). En caso de deficiencias, lo que se recomienda es incorporar 0,5% de Zn en las mezclas minerales que se utilizan como suplemento, especialmente en regiones tropicales y subtropicales (Mufarrege, D. 2000).*

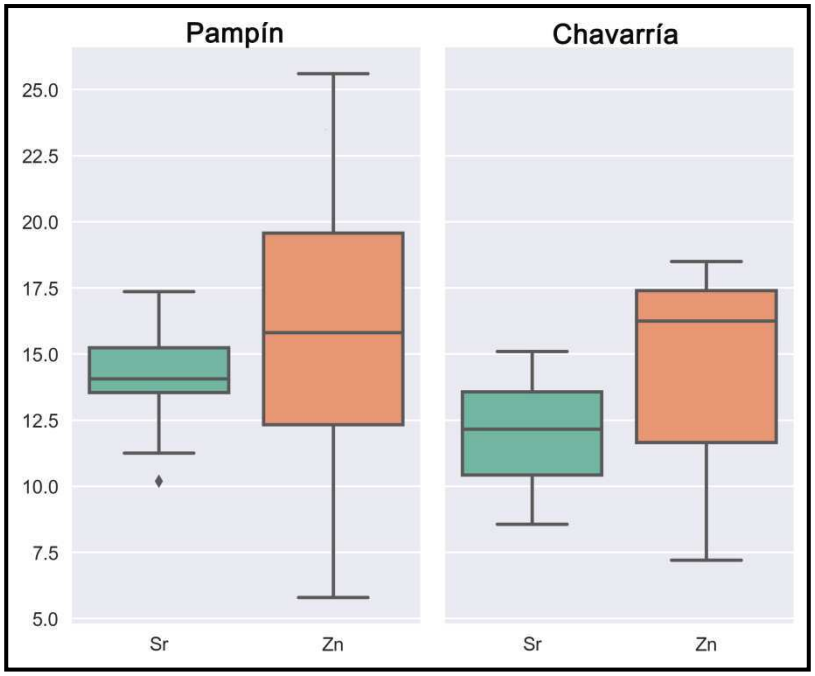


Fig.5.14 Gráficos de Cajas y Bigotes de Sr y Zn en *Andropogon lateralis* de dos series de suelo

5.6 Quimiometría y Aprendizaje Automático

5.6.1 Análisis Exploratorio de datos

En este trabajo se realizó un análisis exploratorio de datos a partir de la composición mineral de las especies forrajeras estudiadas con el objeto de explorar la posible presencia de patrones o relaciones algorítmicas entre las concentraciones de los distintos elementos estudiados. Estos patrones presentes en la estructura de datos pueden brindar información adicional resultante del análisis pormenorizado de los datos con la ayuda de herramientas quimiométricas, en este caso de Análisis Exploratorio de datos. Los objetivos son varios, éstos incluyen: identificar las relaciones entre las variables estudiadas, evidenciar algún resultado con respecto a las hipótesis de trabajo realizadas, controlar que no existan incongruencias en la colección de datos, como datos perdidos, o errores de medida, y la identificación de ciertas áreas donde más datos se necesitan recolectar. Finalmente, el análisis exploratorio de datos es importante porque permite tomar decisiones críticas con respecto a qué camino seguir en el análisis (Brereton, R. G. 2009, Peng, R. D. 2012, Varmuza, K. and Filzmoser, P. 2009).

Para la realización de análisis exploratorio de resultados se aplicaron las siguientes herramientas de manera secuencial:

- Reducción de la dimensionalidad de los datos, mediante un análisis de Componentes Principales (PCA).
- Análisis de las posibles agrupaciones de las variables estudiadas mediante el análisis de conglomerados (HCA).

5.6.2 Análisis de Componentes Principales

La idea principal del análisis de componentes principales es reducir la dimensionalidad de un conjunto de datos de un número largo de variables interrelacionadas, mientras se intenta retener la máxima variación posible presente en los datos. Esto se logra transformando a un nuevo conjunto de variables independientes, que están ordenadas de tal forma que las primeras componentes retienen la mayor variación de las variables originales (Kumar, N. et al. 2014, Varmuza, K. and Filzmoser, P. 2009). Luego de la visualización del análisis de componentes principales, es posible la elección de un modelo de clasificación adecuado (Hackeling, G. 2014).

Para la mejor comprensión del análisis de componentes principales se analizaron tres tipos de gráficos: un gráfico de scores que permite ver la distribución de las muestras en el espacio matemático generado, un gráfico de variables proyectadas en el espacio generado (loadings) que permite observar las posibles correlaciones entre variables y un gráfico de sedimentación en el que se puede ver la contribución relativa de cada componente a la variancia explicada por el modelo.

Los resultados obtenidos al aplicar PCA al contenido mineral en las especies forrajeras se muestran en las Fig. 5.15, 5.16, 5.18, 5.19, 5.21, 5.22.

5.6.3 Análisis de Conglomerados

El análisis de conglomerados es una herramienta –al igual que PCA- de análisis no supervisado que divide el conjunto de datos en conglomerados, o grupos con características similares. Esto lo realiza sin saber antes a que grupo pertenecen estas variables. Las técnicas de agrupamiento se utilizan para tener una visión del

agrupamiento entre los datos. Las técnicas de agrupamiento se guían bajo la premisa de que los objetos dentro de un grupo deben ser más similares entre sí, pero muy diferentes a los objeto de otro eventual grupo. En general, las técnicas de agrupamiento resultan en estructuras que reducen la complejidad y proveen una visión de las relaciones de las variables o muestras estudiadas (Lantz, B. 2015).

Los resultados obtenidos al aplicar un análisis de conglomerados al contenido mineral en las especies forrajeras se muestran en las Fig. 5.17, 5.20 y 5.23.

5.7 Desmodium incanum

5.7.1 Análisis de Componentes Principales

En el gráfico de scores correspondiente a las muestras de *Desmodium incanum* (Fig. 5.15) se observa una distribución de muestras de Pampín con valores positivos en la primera componente principal y dos muestras más alejadas lo que podría eventualmente complicar la posterior clasificación mediante métodos lineales. Las muestras de Chavarría en cambio no están uniformemente distribuidas en el espacio de las dos primeras componentes PC1-PC2.

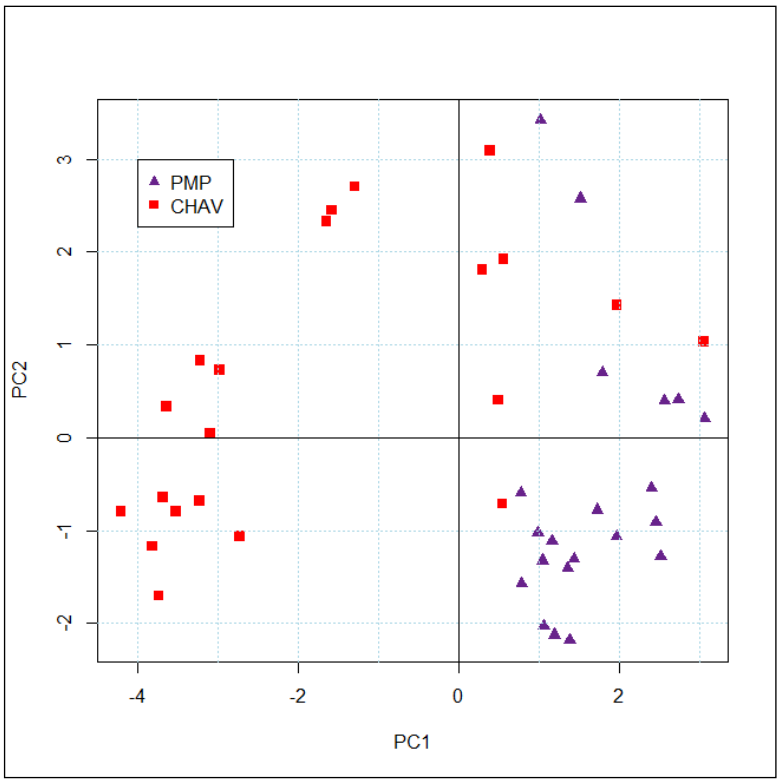


Fig. 5.15 Gráfico de scores correspondiente a muestras de *Desmodium incanum* en las dos series de suelo. **PMP**: Pampín **CHAV**: Chavarría

En el gráfico de loadings (Fig. 5.16) se observa que la primera componentes se vincula de manera positiva con la proyección de vectores de Li, Cu, Sn y Se; y negativamente con la proyección de los vectores de Rb, V y B. La segunda componente se vincula a contribuciones positivas de Zn, Al, Sr y Sb; y contribuciones negativas de Co y Tl. El gráfico de sedimentación muestra contribución de las siguientes componentes con los siguientes valores: PC1: 29,2%; PC2: 12, 6% que se observan de manera análoga en el gráfico de loadings.

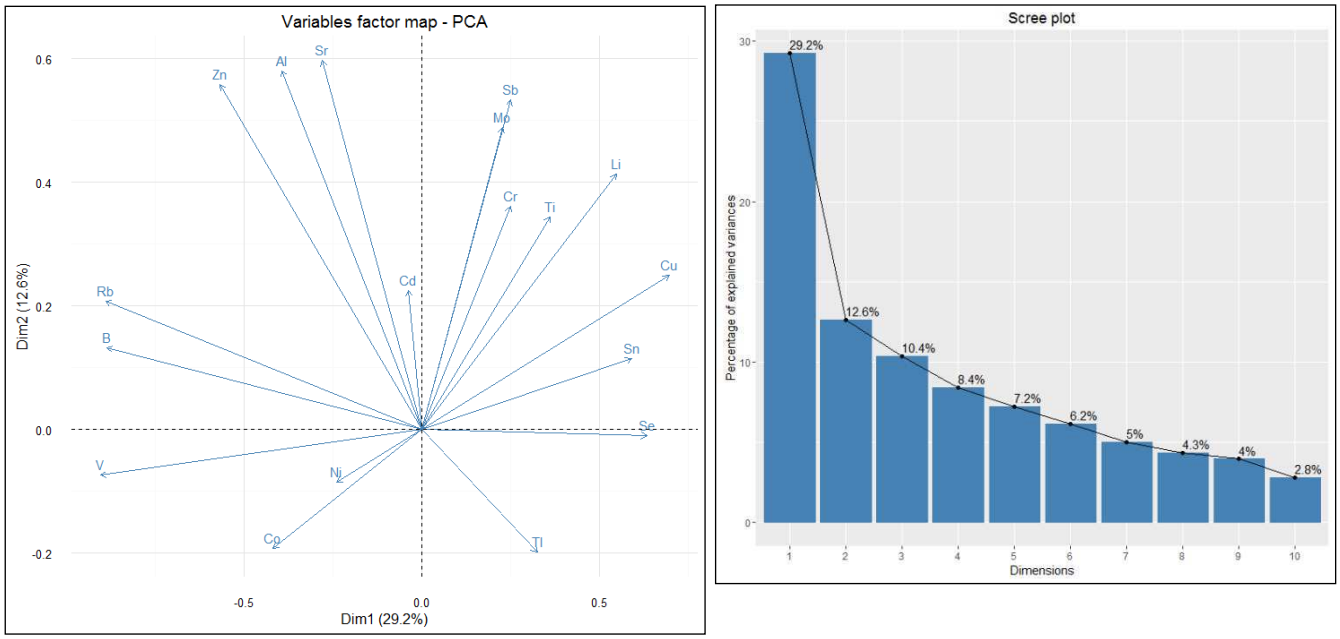


Fig. 5.16 Gráfico de loadings y sedimentación correspondiente a muestras de *Desmodium incanum* en dos series de suelo

5.7.2 Análisis de Conglomerados

La distancia que se utilizó fue de coeficientes de correlación y el método fue el de Ward. En la Fig. 5.17 se observa que las variables están agrupadas en cuatro grupos; que se detallan: Grupo 1: Al, Sr, B, V, Rb, Zn y Co; Grupo 2: Cd y Mo; Grupo 3: Cu, Se, Li, Sn, Ti y Sb; Grupo 4: Ni, Tl y Cr. El dendrograma expuesto es consistente con respecto al gráfico de loadings (Fig. 5.16) de las componentes principales ya que las variables que se encuentran próximas en el gráfico de loadings, a excepción de Ni, Tl y Cr, que se encuentran distribuidas entre las variables y no forman un grupo homogéneo.

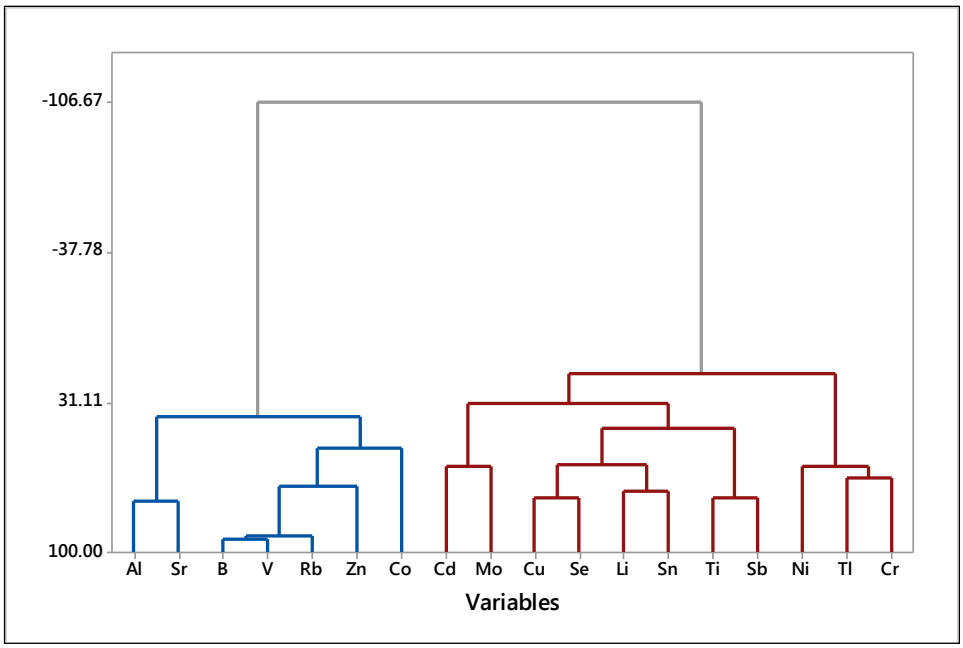


Fig. 5.17 Dendrograma de las variables estudiadas en muestras de *Desmodium incanum* en dos series de suelo

5.8 Schizachyrium microstachyum

5.8.1 Análisis de Componentes Principales

En el gráfico de scores correspondiente a las muestras de *Schizachyrium microstachyum* (Fig. 5.18) se observa una distribución de puntos de Pampín con valores negativos de la primera componente principal, a excepción de una muestra que se encuentra junto a las muestras de Chavarría. Si bien existe un grupo más o menos uniformemente distribuido en muestras de Pampín, cosa que no sucede con las muestras de Chavarría en el gráfico de las componentes PC1-PC2.

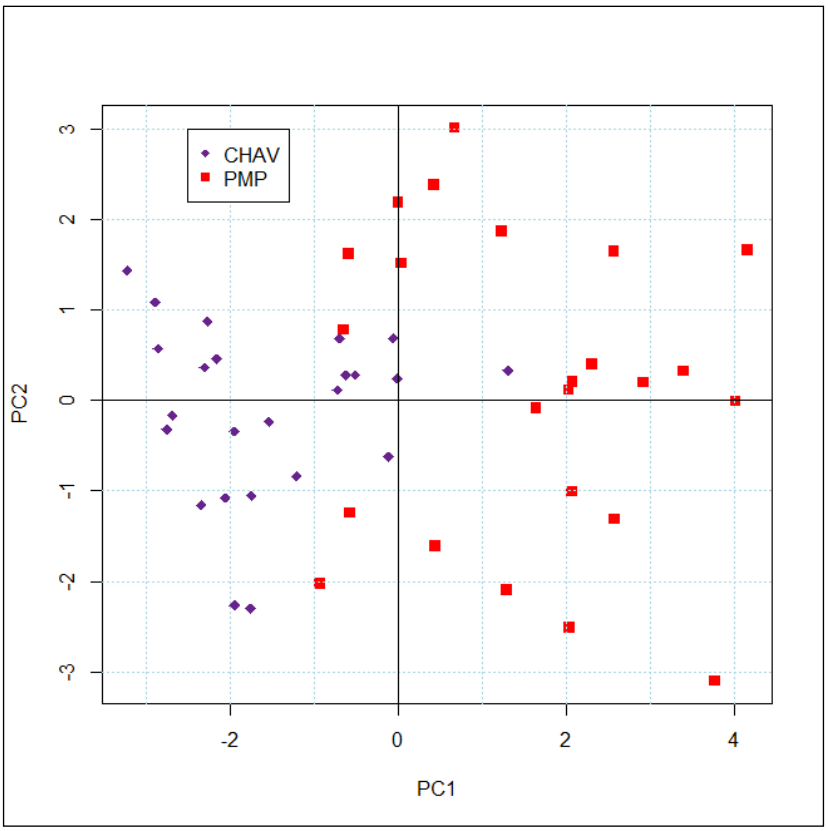


Fig. 5.18 Gráfico de scores de *Schizachyrium microstachyum* en dos series de suelo. **PMP**: Pampín **CHAV**: Chavarría

En el gráfico de loadings (Fig. 5.19) se observa que la primera componente (PC1) se vinculan de manera positiva con la proyección de vectores de Tl, Mo, Li, Cu y Sn; y negativamente con la proyección de los vectores de Ni, Sr y V. La segunda componente se vincula a contribuciones positivas de Ni y Tl; y contribuciones negativas de B, Ti y Sn. El gráfico de sedimentación muestra contribución de las siguientes componentes con los siguientes valores: PC1: 23%; PC2: 10, 1% que se observan de manera análoga en el gráfico de loadings.

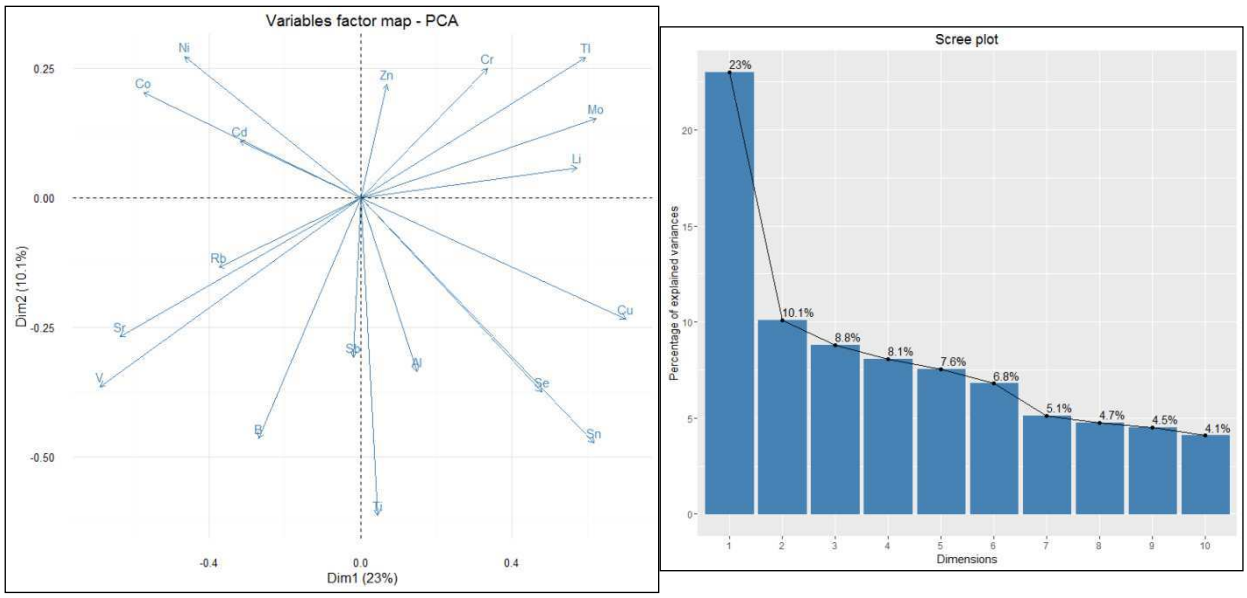


Fig. 5.19 Gráfico de loadings y sedimentación de *Schizachyrium microstachyum* en dos series de suelo

5.8.2 Análisis de Conglomerados

La distancia que se utilizó fue de coeficientes de correlación y el método fue el de Ward. En la Fig. 5.20 se observa que las variables están agrupadas en cuatro grupos; que se detallan: Grupo 1: Al, Cu, Se, Sn, Li, Tl, Mo y Cr; Grupo 2: B, Rb y Ti; Grupo 3: Zn y Sb; Grupo 4: Cd, Sr, V, Co y Ni. El dendrograma expuesto es consistente con respecto al gráfico de loadings (Fig. 5.19) de las componentes principales con respecto al Grupo 1 de variables. El Grupo 2 y el 3 de variables no forman un grupo homogéneo de variables ya que se encuentran distribuidas no uniformemente en el gráfico de loadings. El grupo 4 forma un grupo más homogéneo a diferencia de los anteriores, sin contar el Rb que aparece próximo en Fig. 5.19 y no así en el dendrograma.

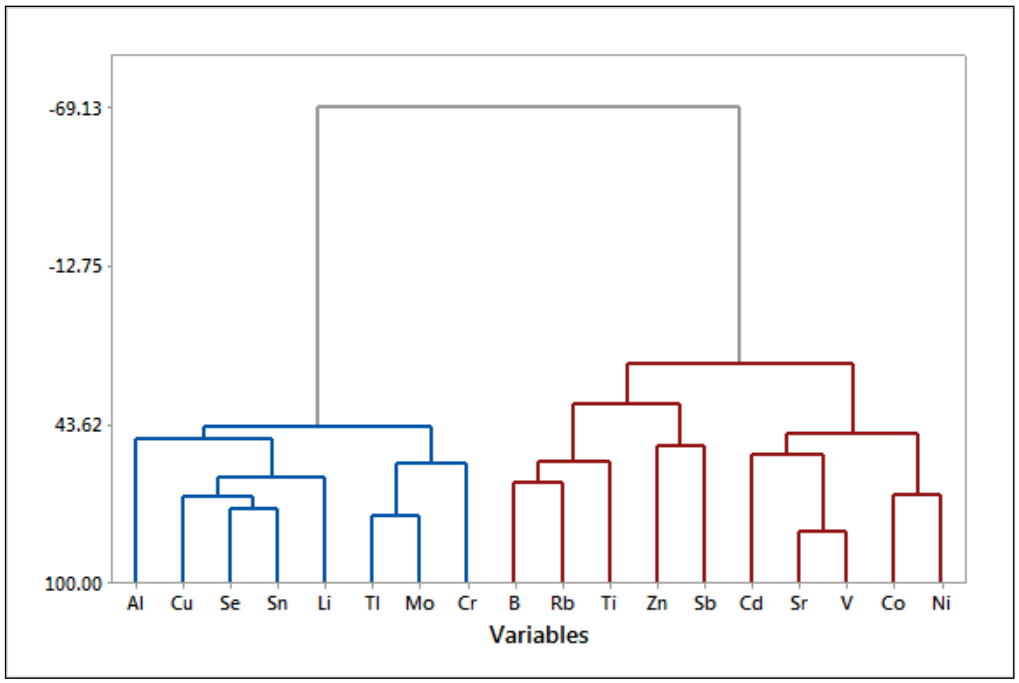


Fig. 5.20 Dendrograma de las variables estudiadas en *Schizachyrium microstachyum* muestras de en dos series de suelo

5.9 Andropogon lateralis

5.9.1 Análisis de Componentes Principales

De la proyección de muestras en el espacio de las componentes principales generadas se puede observar la presencia de cierta tendencia a formar grupos, lo que indicaría cierta similitud entre las muestras estudiadas, a pesar de que la PCA no tiene como uno de sus objetivos clasificar muestras, sino ver la distribución de muestras en el espacio matemático generado intentando representar la mayor proporción de información presente de los datos en las dos primeras componentes generadas.

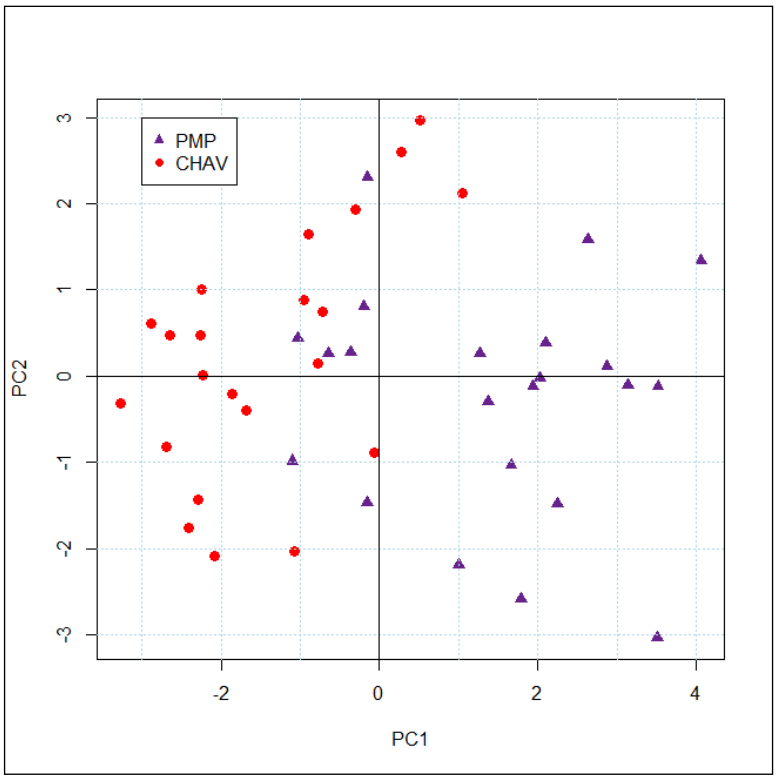


Fig. 5.21 Gráfico de scores de la composición mineral de *Andropogon lateralis* en las dos series de suelo. **PMP**: Pampín **CHAV**: Chavarría

Específicamente en la Fig. 5.21 de muestras de *Andropogon lateralis*, se ven agrupamientos de muestras de Chavarría y Pampín, aunque este agrupamiento no es nítido, lo cual hará necesario la aplicación de métodos subsiguientes –no lineales- con el fin de clasificar las muestras según serie de suelo. En el gráfico de loadings (Fig. 5.22) se observa que las componentes se vinculan de manera positiva con la proyección de vectores de Sn, Cu, Li, Mo y Tl; y negativamente con la proyección de los vectores de B, Sr, Co y Ni entre otros. La segunda componente se vincula a valores altos de Ti, Sn, Tl y Ni entre otros. A modo ilustrativo se reproduce un gráfico de sedimentación a la derecha para ver la contribución de las diferentes componentes. De modo práctico en este trabajo se eligen las dos primeras componentes para su visualización debido a su claridad.

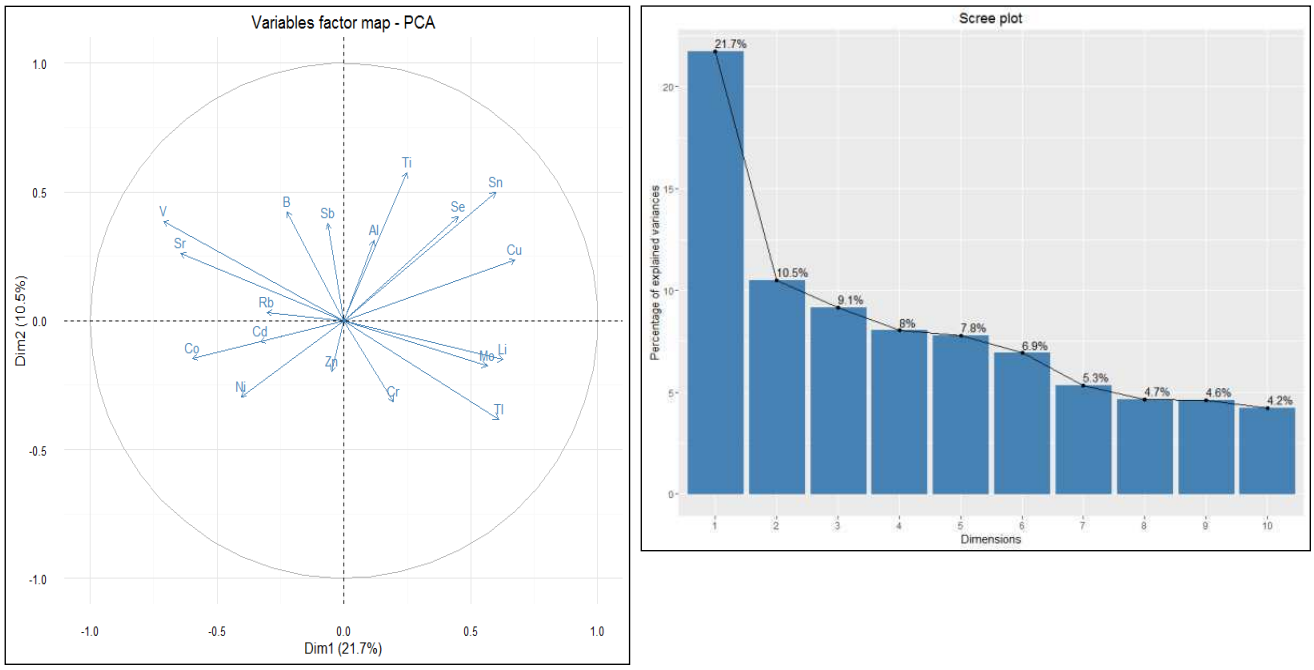


Fig. 5.22 Gráfico de loadings y de sedimentación de la composición mineral de *Andropogon lateralis*

5.9.2 Análisis de Conglomerados

La distancia que se utilizó fue de coeficientes de correlación y el método fue el de Ward. En la Fig. 5.23 se observa que las variables están agrupadas en cuatro grupos; que se detallan: Grupo 1: Al, Cu, Se, Sn, Li y Ti; Grupo 2: Tl, Mo y Cr; Grupo 3: B, Rb, Zn y Sb; Grupo 4: Cd, Sr, V, Co y Ni. El dendrograma expuesto es consistente con respecto al gráfico de loadings (Fig. 5.22) de las componentes principales ya que las variables que se encuentran próximas en el gráfico de loadings, también lo están en el dendrograma.

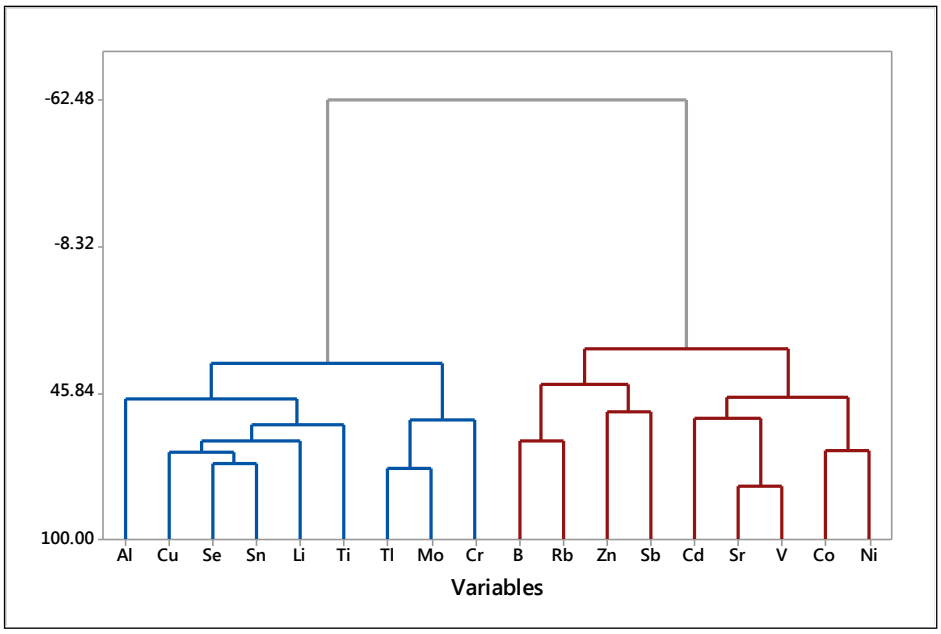


Fig. 5.23 Dendrograma de las variables estudiadas en muestras de *Andropogon lateralis* en dos series de suelo

5.10 Propuesta de Modelos de Aprendizaje Automático para Clasificación de Forrajes

En las siguientes secciones se proponen modelos para lograr una clasificación con valores de exactitud aceptables. La complejidad de la función de clasificación está íntimamente relacionada a la complejidad del conjunto de datos, y esto se observa a lo largo de los pasos que se emplearon para clasificar los diferentes conjuntos de datos pertenecientes a las especies forrajeras. En todos los casos estamos frente a una clasificación binaria de datos con conjuntos de datos balanceados, es decir, el mismo número de muestras en cada grupo, lo que nos permite tomar la exactitud como un parámetro adecuado como criterio de bondad.

5.10.1 Desmodium incanum

5.10.1.1 Análisis Discriminante Lineal (LDA)

La técnica de análisis discriminante lineal es una técnica de análisis supervisado utilizada principalmente para clasificar muestras de acuerdo a dos o más grupos, de acuerdo a un conjunto de variables que describen el grupo de datos, en este, la concentración mineral (Kruzlicova, D. *et al*. 2013). Es frecuente encontrar en la bibliografía que es presentada como una técnica de reducción de dimensión –al igual que el análisis de componentes lineales-, aunque la mayor diferencia con ésta en maximizar la variancia inter-grupos, mientras se disminuye la variancia intra-grupos. Mientras que el análisis de componentes principales intenta encontrar los ejes ortogonales de máxima variancia en el set de datos; el objetivo en el análisis discriminante lineal es encontrar el subespacio que optimiza la separación de esas clases (Raschka, S. 2015). En la Fig. 5.24 se graficaron las muestras de *Desmodium incanum* según las dos funciones canónicas que fabrica el modelo para representar en un espacio matemático las muestras estudiadas. Se observan dos grupos, las muestras de Chavarría con valores positivos de la segunda función canónica, y las muestras de Pampín con valores negativos, a excepción de una sola muestra que fue proyectada en la zona de valores positivos de la función.

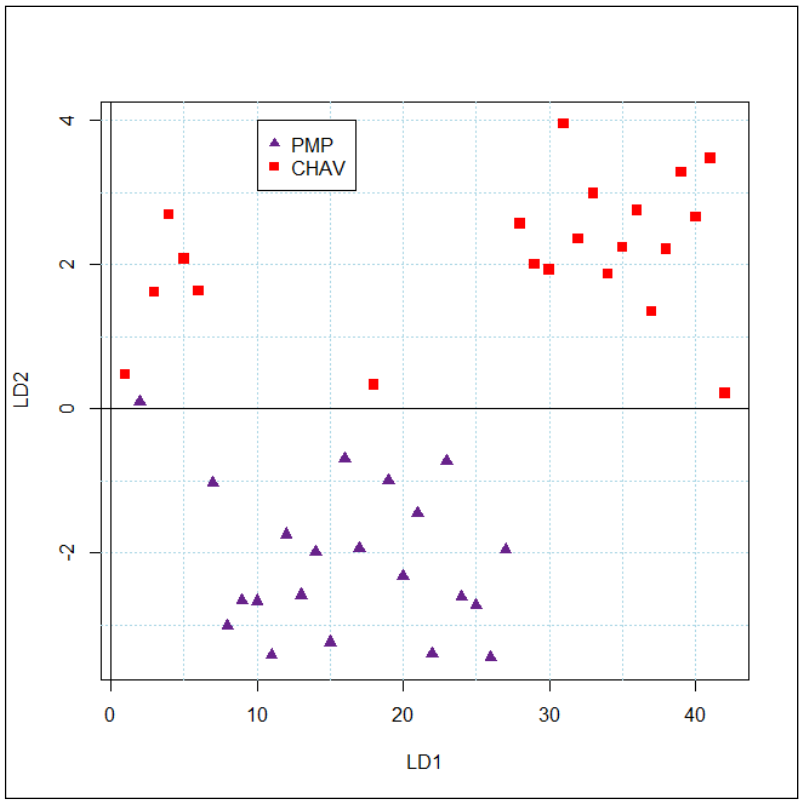


Fig. 5.24 Análisis discriminante de composición mineral de *Desmodium incanum* en dos series de suelo. **PMP**: Pampín **CHAV**: Chavarría

En una segunda etapa se analizó la matriz de confusión para observar las métricas de interés que se obtuvieron con análisis discriminante lineal, para los dos sets de datos estudiados: datos de entrenamiento y datos de prueba (Kirk, M. 2014, Raschka, S. 2015). Existen varias maneras de dividir un set de datos:

- Método de Retención (Hold-out): asumiendo que todas las muestras son independientes y están distribuidas de manera idéntica, se toma una parte del set de datos para la construcción del modelo y luego el set restante se utiliza para la validación. Computacionalmente hablando, el método hold-out es simple de programar y fácil de correr. Una de las desventajas es su menor poder estadístico. Como los resultados de validación provienen de un set de datos pequeño el error de generalización es menos confiable. En general se

utiliza este método ante set de datos grandes, de tal modo que el set de validación sea lo suficientemente grande para asegurar estimaciones estadísticas confiables.

- Método de validación cruzada (Cross-validation): El método de validación cruzada permite dividir el set de datos en *k*-secciones, entrenar en k-1 sets de datos y evaluarlo en el *k* restante.
- Método de remuestreo (Bootstrap resampling): el bootstrap es una técnica de remuestreo y se caracteriza por generar múltiples sets de datos. Cada uno de los sets de datos se utiliza para estimar alguna cantidad de interés. Como hay múltiples sets de datos y múltiples estimadores, uno puede también calcular un intervalo de confianza para el estimador. A diferencia de los métodos mencionados previamente; el bootstrap es un método de remuestreo con reemplazo.

Estas tres diferentes metodologías para la división de un conjunto de datos se observan esquemáticamente en la Fig. 5.25

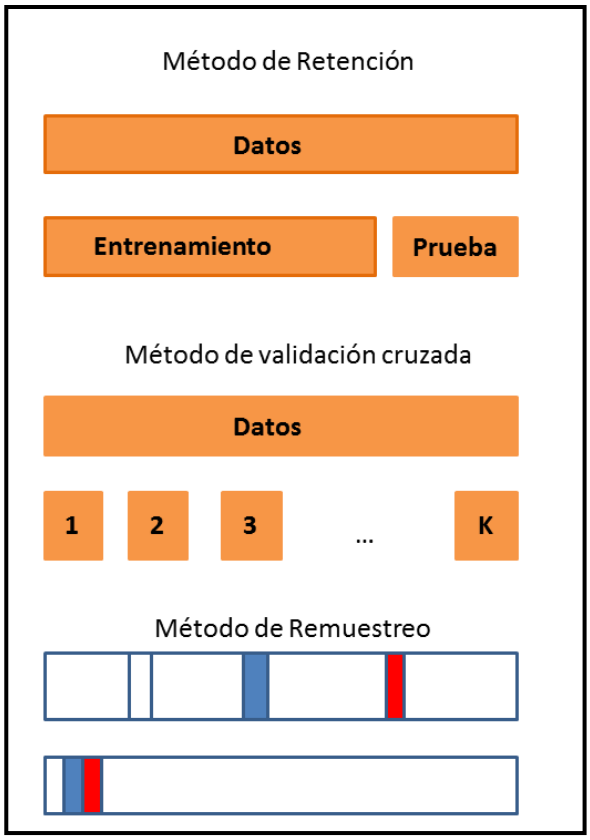


Fig. 5.25 Esquema correspondiente a diferentes procedimientos de remuestreo

En este caso en particular se utilizó el método de *k*-validación cruzada. A continuación (Fig. 5.26) se muestra un esquema para la mejor comprensión del método de *k*-validación cruzada dentro del esquema de trabajo. Dividir un set de datos para obtener los hiper-parámetros de un modelo y luego evaluarlo es una de las formas de combatir el sobreajuste en herramientas de análisis predictivo o de aprendizaje automático, la otra es la regularización (Abu-Mostafa, Y. S. *et al*. 2012), pero será explicada cuando sea necesaria la búsqueda de hiper-parámetros óptimos.

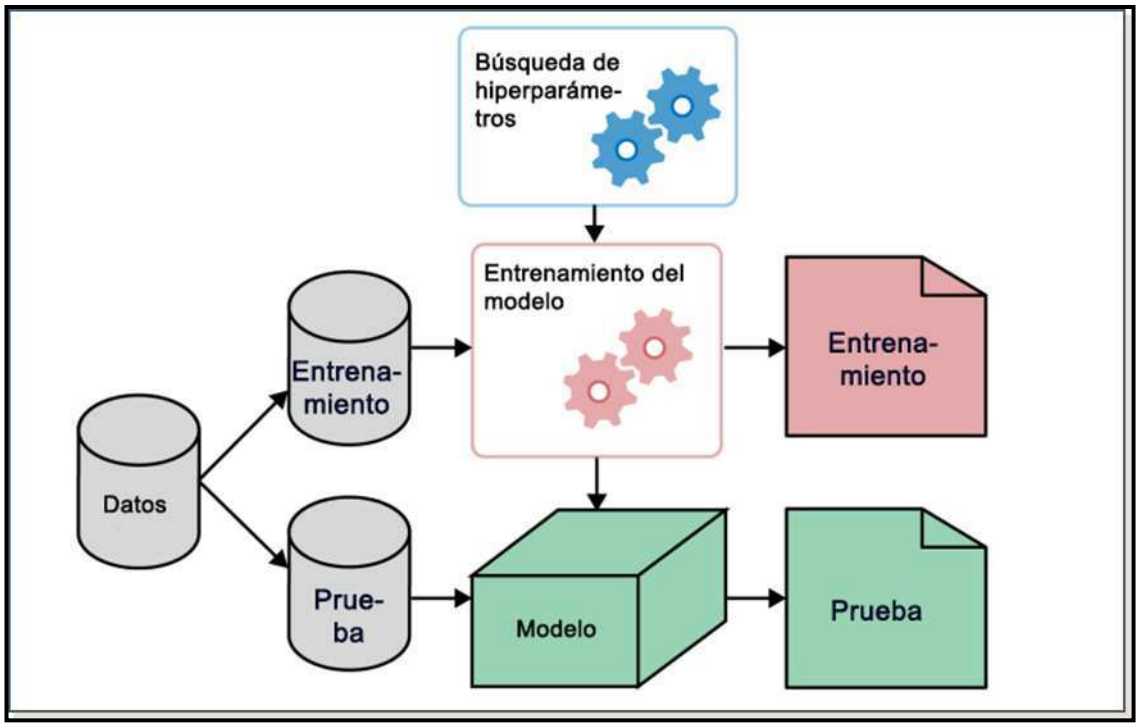


Fig. 5.26 Esquema de trabajo

La matriz de confusión nos arroja valores que son de interés para evaluar el desempeño de un clasificador en particular; a saber (Japkowicz, N. and Shah, M. 2011):

- Exactitud (ACC):

$$ACC = \frac{VP + VN}{VP + VN + FP + FN}$$

- Sensibilidad o tasa de verdaderos positivos (REC):

$$REC = TPR = \frac{VP}{P} = \frac{VP}{FN + VP}$$

- Especificidad o tasa de verdaderos negativos (SPEC):

$$SPEC = \frac{VN}{N} = \frac{VN}{VN + FP}$$

- Precisión o tasa de predichos positivos (PRE):

$$PRE = \frac{VP}{VP + FP}$$

Se presenta en la Tabla 5.5 la matriz de confusión correspondiente para evaluar las métricas correspondientes a *Desmodium incanum.* Para la etapa de entrenamiento del modelo se utilizaron en total 32 muestras y, para la prueba se utilizaron 10 muestras. En la Tabla 5.5 se expresan los valores porcentuales de la matriz de confusión y las métricas correspondientes también en valores porcentuales.

Tabla 5.5 Matriz de confusión de Análisis discriminante lineal para *Desmodium incanum*

	Chavarría	Pampín
Chavarría	100%	0
Pampín	14%	86%
Exactitud: 90%		
Especificidad (TVN): 86%		
Sensibilidad (TVP): 100%		
Precisión: 75%		
AUC: 90%		

Al analizar la matriz de confusión en valores absolutos se puede mencionar que se cometió un solo error en la clasificación de muestras de Pampín, lo que se expresa en un valor de 14% de los falsos positivos (FP), y un 100% de los casos verdaderos positivos que se expresan mediante la sensibilidad.

Se observa una clasificación óptima en el caso de las muestras de Chavarría, lo que luego se traduce en un valor alto de sensibilidad. El valor de exactitud obtenido también es un valor aceptable para los fines establecidos, lo que hace innecesario seguir en el análisis de algún otro algoritmo con desempeño mejor que LDA.

También para evaluar el desempeño de un algoritmo de aprendizaje automático, existen métodos gráficos (Japkowicz, N. and Shah, M. 2011). Entre ellos podemos mencionar de interés para este trabajo:

- Curvas ROC: una curva ROC es un gráfico que en el eje horizontal grafica la tasa de falsos positivos (TFP), y el eje vertical la tasa de los verdaderos positivos (TVP) de un clasificador en particular. En otras palabras, podemos decir que TVP es la sensibilidad mientras que TFP es 1- especificidad de dicho modelo estudiado (Japkowicz, N. and Shah, M. 2011, Raschka, S. 2015). Un valor que nos resulta de interés a la hora de evaluar una curva ROC es el área bajo la curva o AUC, que es mejor cuanto más se acerque a uno. La línea punteada diagonal puede ser interpretada como un resultado aleatorio, y un modelo de clasificación que esté por debajo de esta línea, es considerado con un desempeño más bajo que un resultado aleatorio. Un clasificador óptimo caería en el extremo superior izquierdo con un valor de TVP igual a 1, y un TFP de 0.
- Curvas de precisión – sensibilidad (curvas PR): estas curvas se asemejan a las curvas ROC en el sentido de que exploran el equilibrio entre los casos positivos bien clasificados y el número de casos negativos mal clasificados. Como el nombre sugiere, una curva de Precisión – Sensibilidad lo que hace es graficar la Precisión en función de la Sensibilidad, o, en otras palabras, evalúa la distinta precisión que se puede obtener a diferentes sensibilidades. A diferencia de una curva ROC, una curva PR tiene una pendiente negativa.

A continuación, se presentan las curvas correspondientes (ROC y PR) para *Desmodium incanum* con Análisis discriminante lineal (Fig. 5.27 y Fig. 5.28). Los valores altos obtenidos para la matriz de confusión, hace que las curvas ROC y PR obtenidas sean cercanas a las ideales para un clasificador binario, esto se observa también en el valor alto del área bajo la curva que fue de 0,95.

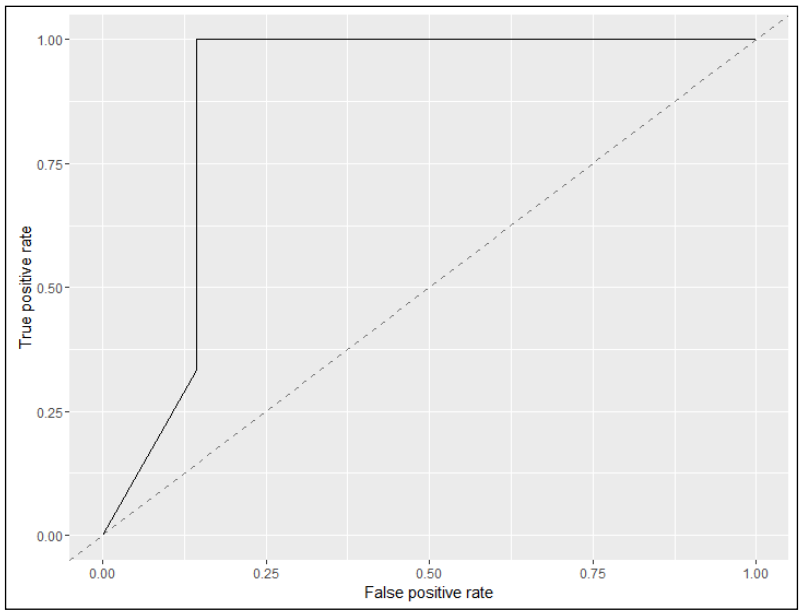


Fig. 5.27 Curva ROC correspondiente a *Desmodium incanum* con análisis discriminante lineal

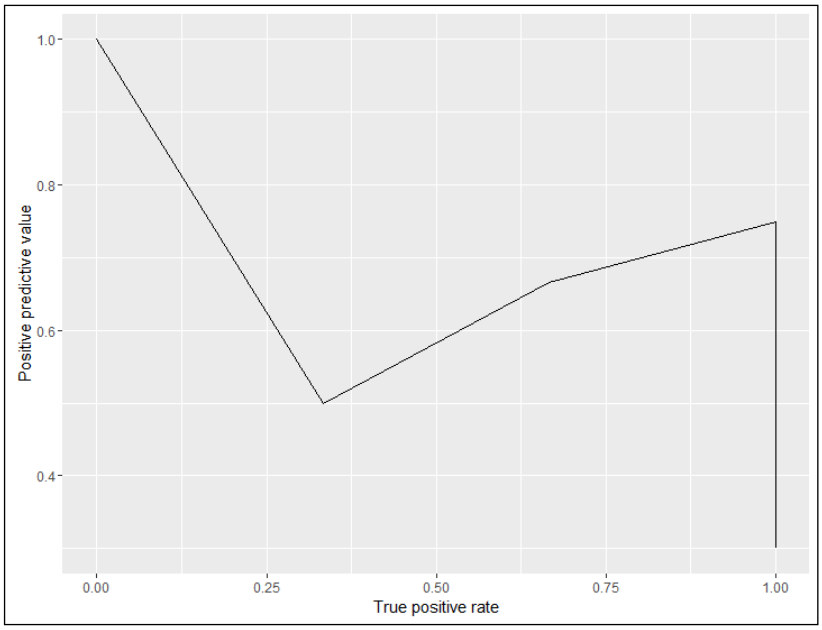


Fig. 5.28 Curva PR correspondiente a Desmodium incanum con análisis discriminante lineal

5.10.2 Schizachyrium microstachyum

5.10.2.1 Análisis Discriminante Lineal (LDA)

En una primera etapa de clasificación se realizó un gráfico de análisis discriminante lineal y se evaluaron los resultados de interés que nos arroja la matriz de confusión.

En la Fig. 5.29 se observa una distribución de las muestras de Pampín en el espacio positivo de la función discriminante lineal (LD2), a excepción de una muestra que aparece proyectada en el espacio negativo y otra muestra que aparece a una distancia mayor que el grupo de muestras de Pampín. El grupo de muestras de Chavarría aparece en el espacio negativo de la proyección de la función LD2, a excepción de dos muestras que aparecen junto a las de Pampín.

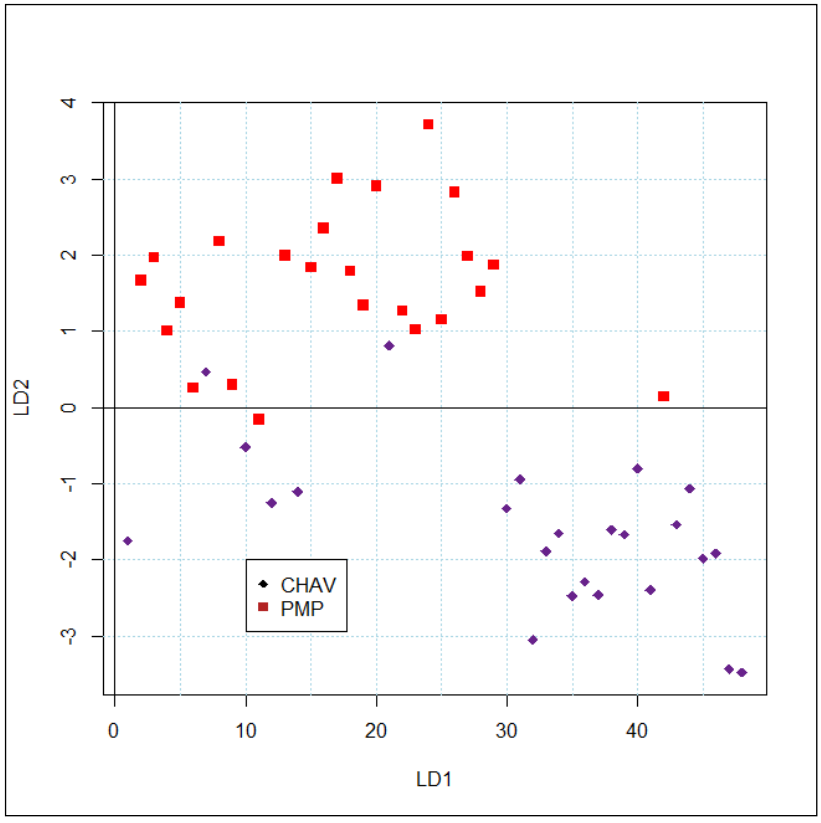


Fig. 5.29 Análisis discriminante de composición mineral de *Schizachyrium microstachyum* en las dos series de suelo. **PMP**: Pampín **CHAV**: Chavarría

A continuación, se presenta la matriz de confusión (Tabla 5.6) con las métricas de interés para la clasificación. Para la etapa de entrenamiento del modelo se emplearon en total 36 muestras, y finalmente 12 muestras para la etapa de prueba. En la Tabla 5.6 se expresan los valores porcentuales de la matriz de confusión y de las métricas correspondientes.

Tabla 5.6 Matriz de confusión de *Schizachyrium microstachyum* con LDA como clasificador

	Chavarría	Pampín
Chavarría	83%	17%
Pampín	50%	50%
Exactitud: 67%		
Especificidad (TVN): 50%		
Sensibilidad (TVP): 83%		
Precisión: 62%		
AUC = 63%		

Al observar la matriz de confusión de los valores absolutos, se observa un error en las muestras de Chavarría que fue clasificado como Pampín (17%) y tres muestras de Pampín clasificadas como Chavarría, de un total de seis muestras totales, lo que se expresa en valores de 50% tanto para Chavarría como Pampín. Se observaron cuatro errores totales de las 12 muestras totales utilizadas como prueba.

Si se observa atentamente a la distribución de puntos se observa que no están distribuidos de manera homogénea, lo que complica la clasificación mediante un método lineal, ya que las funciones propuestas van a ser de una línea recta. A esto podemos atribuir el bajo valor de exactitud obtenido por LDA. Se observa en la curva ROC correspondiente que, si bien se acerca a la zona superior izquierda, aún está lejos de ser el clasificador óptimo para el set de datos. El desempeño cae de igual forma en

la curva PR ya que el valor global de precisión y sensibilidad rondan alrededor de 62% y 83% respectivamente.

A continuación, se presenta la curva ROC correspondiente (Fig. 5.30) y la curva PR (Fig. 5.31) correspondiente.

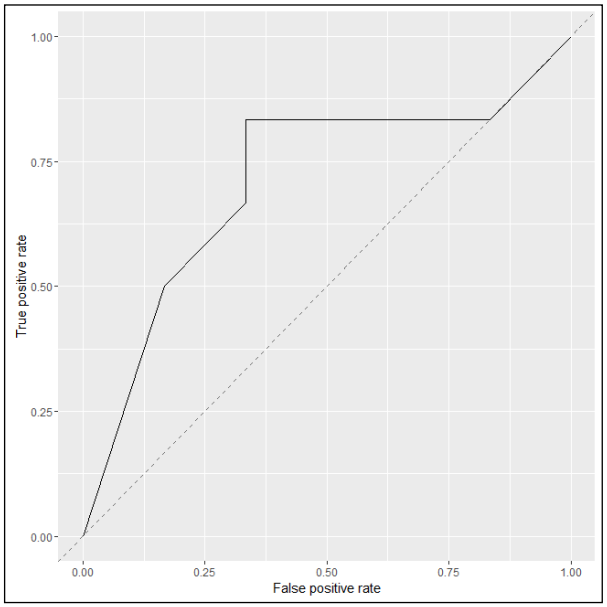


Fig. 5.30 Curva ROC correspondiente a *Schizachyrium microstachyum* con análisis discriminante lineal

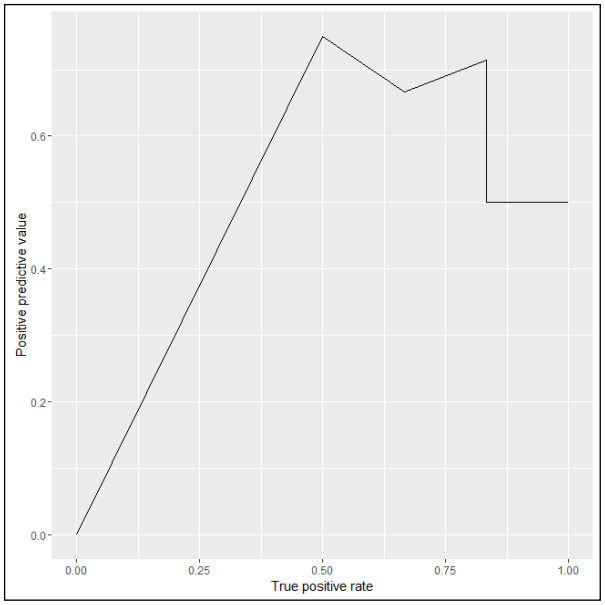


Fig. 5.31 Curva PR correspondiente a *Schizachyrium microstachyum* con análisis discriminante lineal

Como se ha observado previamente la distribución no homogénea de muestras en un espacio lineal hace necesario considerar el empleo de clasificadores no lineales en el siguiente paso para obtener mejores valores de clasificación (porcentajes de exactitud mayores).

5.10.2.2 Support Vector Machines

Support Vector Machines (SVM) es el nombre de una técnica de clasificación supervisada de datos, en la que se proyectan las muestras en estudio en nuevo espacio multidimensional creado por una función. La técnica de Máquinas de Soporte Vectoriales es en esencia una técnica lineal de clasificación debido a que el límite de decisión que propone es un hiperplano (Lantz, B. 2015) pero gracias al truco del "kernel" se convierte en una potente herramienta de clasificación para muestras que presentan comportamiento no lineal (Forte, R. M. 2015, Hackeling, G. 2014). SVM utiliza como técnica de pre-tratamiento de datos, el escalado ya que esto facilita el aprendizaje de parámetros internos para la correcta clasificación (Raschka, S. 2015).

El siguiente paso consiste en optimizar hiper-parámetros que son los coeficientes de la función propuesta para el kernel de SVM. En el caso del kernel radial propuesto; son dos: C y σ. El valor de costo o C se aplica a todos los modelos que violan las restricciones, y en vez de encontrar el máximo margen, el algoritmo intenta llevar a un mínimo ese valor. Un valor alto de C intentará encontrar un 100% de separación. Sin embargo, un valor bajo de C pondrá el énfasis en un margen en general más largo

(Lantz, B. 2015). Otro hiper-parámetro a optimizar es el σ, que se encuentra en el numerador de la función radial:

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = e^{\frac{-\|\overrightarrow{x_i}-\overrightarrow{x_j}\|}{2\sigma^2}}$$

El parámetro σ, indica el diámetro de la función gaussiana empleada en el kernel utilizado para los límites de decisión. Se puede observar la relación en la Fig 5.32

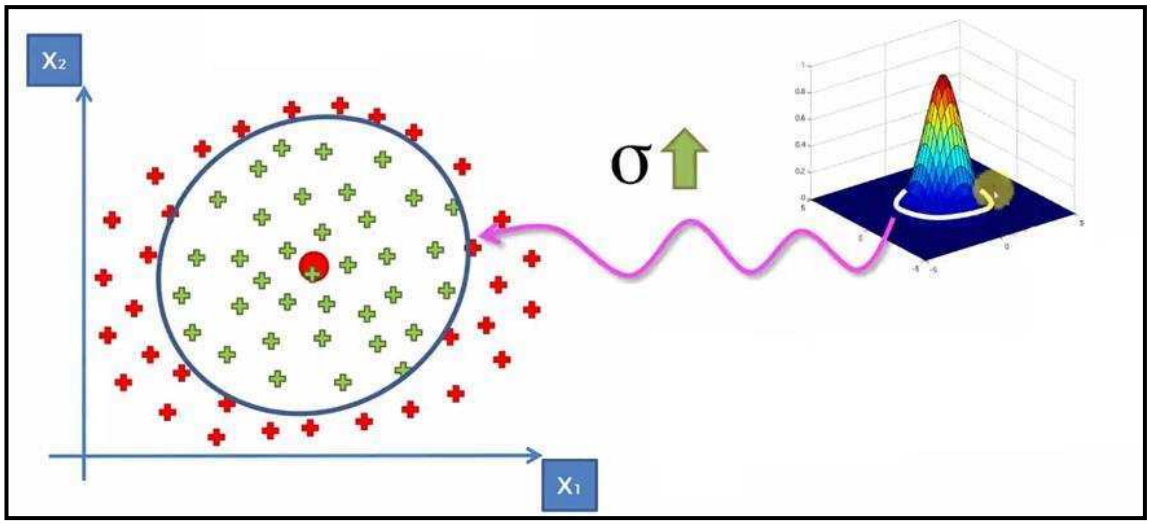


Fig 5.32 Variación de σ y su relación con el límite de decisión en SVM

Una de las maneras de combatir el sobreajuste en SVM es mediante la optimización de hiper-parámetros óptimos del kernel; con una validación cruzada de los datos, como se vino trabajando anteriormente. Para optimizar los valores de C y σ, se puede hacer una búsqueda mediante un grid o una búsqueda aleatoria. En el primer caso se debe especificar los valores a probar por la función para encontrar los valores óptimos y en el segundo caso se deben especificar los límites inferior y superior ya que la búsqueda aleatoria es continua. De las dos, la búsqueda mediante un grid es computacionalmente más costosa. En la Fig 5.33 se observan las diferencias entre una

búsqueda mediante un grid (grid search) y una búsqueda aleatoria (random search) (Goodfellow, I. et al. 2016).

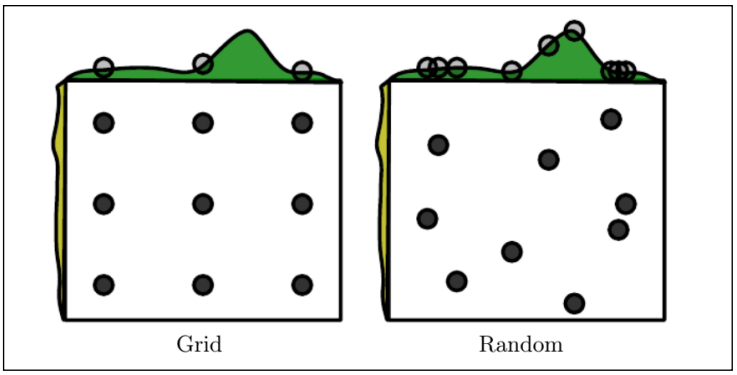


Fig 5.33 Diferencias entre una búsqueda grid y una búsqueda aleatoria para optimización de hiper-parámetros

A continuación, se muestra la imagen correspondiente a un mapeo de posibles valores de C y sigma, y los posibles valores exactitud en forma gráfica correspondiente al set de testeo (Fig 5.34).

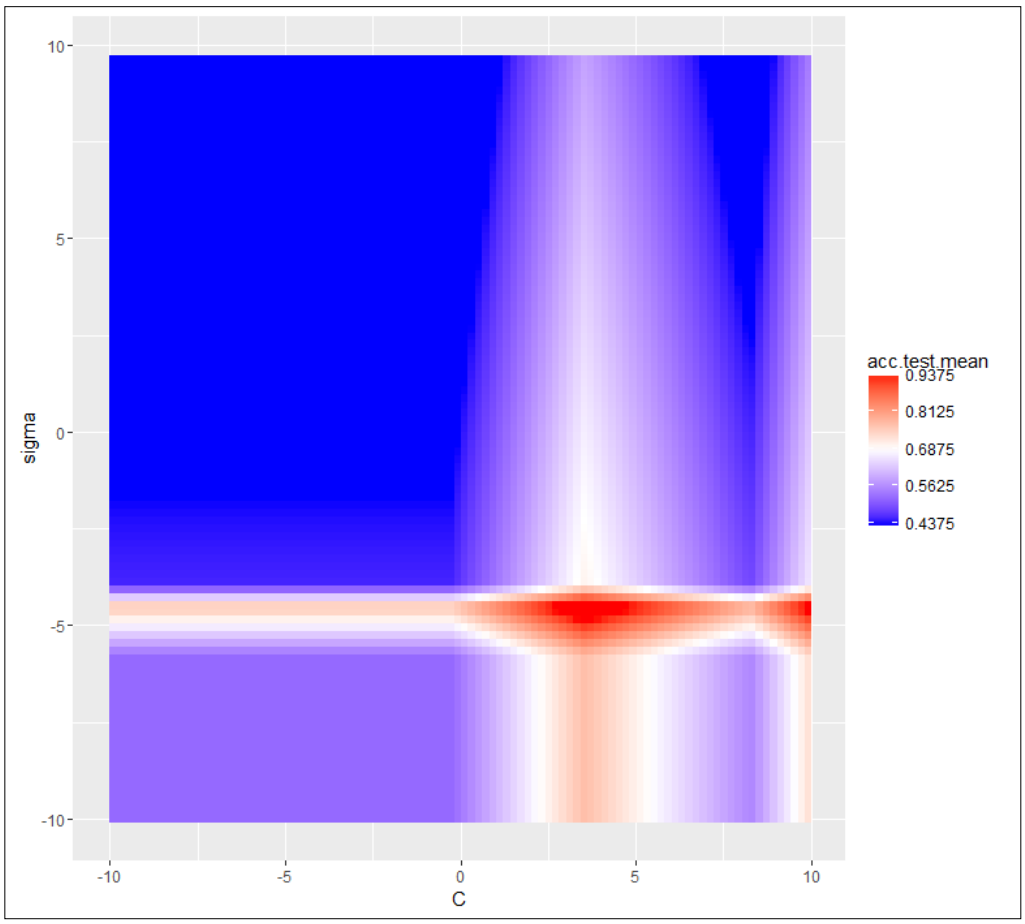


Fig. 5.34 Mapeo de C y sigma en búsqueda de los valores óptimos

Una vez ya con C y σ definidos (C: 0,501 y σ: 0,101), vamos a obtener el modelo SVM con el set de entrenamiento, y evaluar el desempeño en el conjunto de datos de prueba.

A continuación, en la Tabla 5.7, se presenta los valores de la matriz de confusión para evaluar el desempeño del algoritmo de clasificación.

Tabla 5.7 Matriz de confusión de Support Vector Machines para *Schizachyrium microstachyum*

	Chavarría	Pampín
Chavarría	83%	17%
Pampín	0%	100%
Exactitud: 92%		
Especificidad (TVN): 100%		
Sensibilidad (TVP): 83%		
Precisión: 100%		

AUC: 91.6%

Para la confección de la matriz de confusión se utilizaron 36 muestras para la etapa de entrenamiento y 12 muestras para la etapa de prueba. La matriz de confusión de valores absolutos expresa un solo error en las muestras de Chavarría predicha como Pampín, y las muestras de Pampín fueron correctamente clasificadas en su totalidad. El número total de errores en valores absolutos es de una sola muestra, lo que se expresa en un porcentaje de clasificación mayor a 90%

Los resultados de la matriz de confusión dan valores óptimos de clasificación, y sólo hay errores en las muestras de Chavarría. Los valores obtenidos son superiores a los obtenidos con un método lineal como LDA, lo que nos permite afirmar que SVM es un mejor clasificador al comparar las métricas correspondientes.

En la Fig. 5.35 se observa como los valores ideales de sensibilidad o TPR de 100% y de especificidad nos da como resultado una curva ROC que se corresponde al clasificador ideal. Acá se observa que a pesar de que tener un gráfico ideal, el valor de exactitud es un poco menor de 92%, este valor no está reflejado en el gráfico ROC. La curva PR (Fig. 5.34) obtenida también es la ideal, debido a los valores altos que se obtuvieron en las métricas de Precisión (100%) y Sensibilidad (89%).

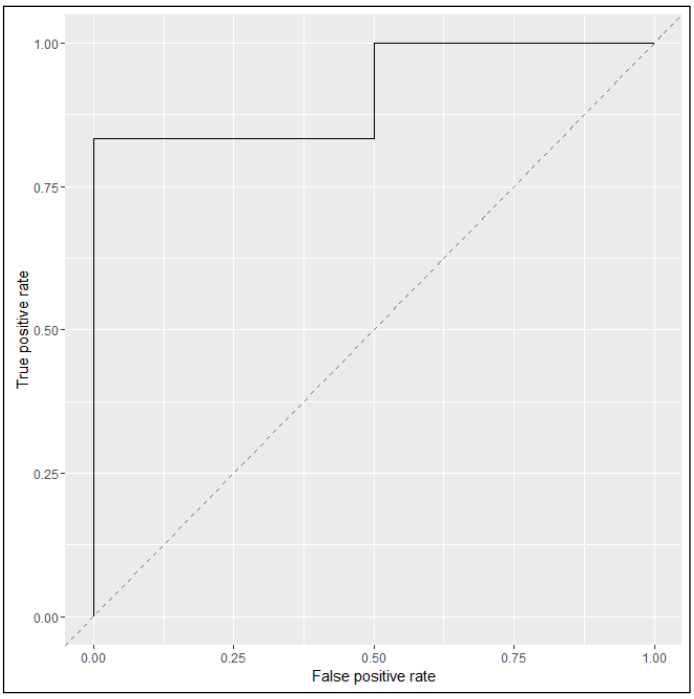


Fig. 5.35 Curva ROC correspondiente a *Schizachyrium microstachyum* con support vector machines

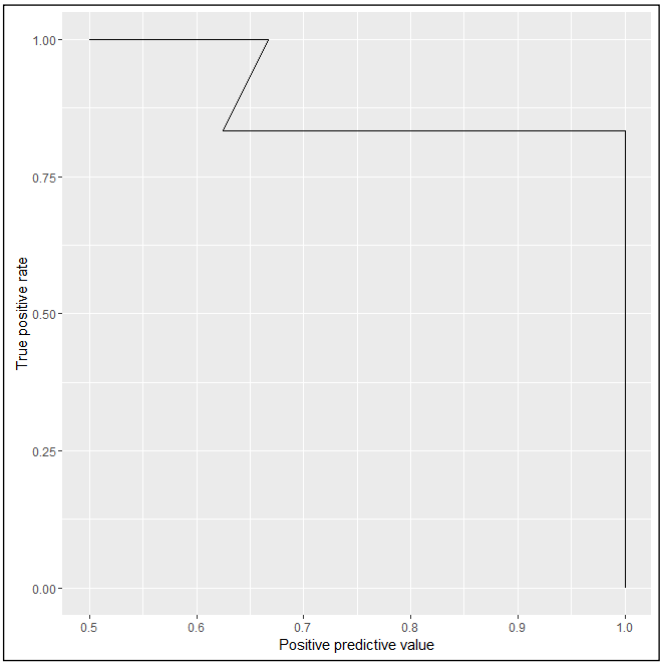


Fig. 5.36. Curva PR correspondiente a *Schizachyrium microstachyum* con support vector machines

5.10.3 Andropogon lateralis

5.10.3.1 Análisis Discriminante Lineal (LDA)

En una primera etapa se realizó un análisis discriminante lineal para ver el desempeño en la clasificación de las muestras de *Andropogon lateralis* según las series de suelo Chavarría y Pampín. Todas las muestras están proyectadas en el espacio dimensional positivo de la función discriminante lineal (LD1), pero la diferencia entre Chavarría y Pampín se observa en espacios diferentes. Las muestras de Chavarría están en un espacio positivo, a diferencia de dos muestras, una proyectada en el espacio negativo de la segunda función (LD2) y otra sobre el eje. Las muestras de Pampín están en un espacio dimensional negativo de la función LD2, a excepción de una muestra que se encuentra junto a las muestras de Chavarría. El gráfico correspondiente se presenta en la Fig. 5.37

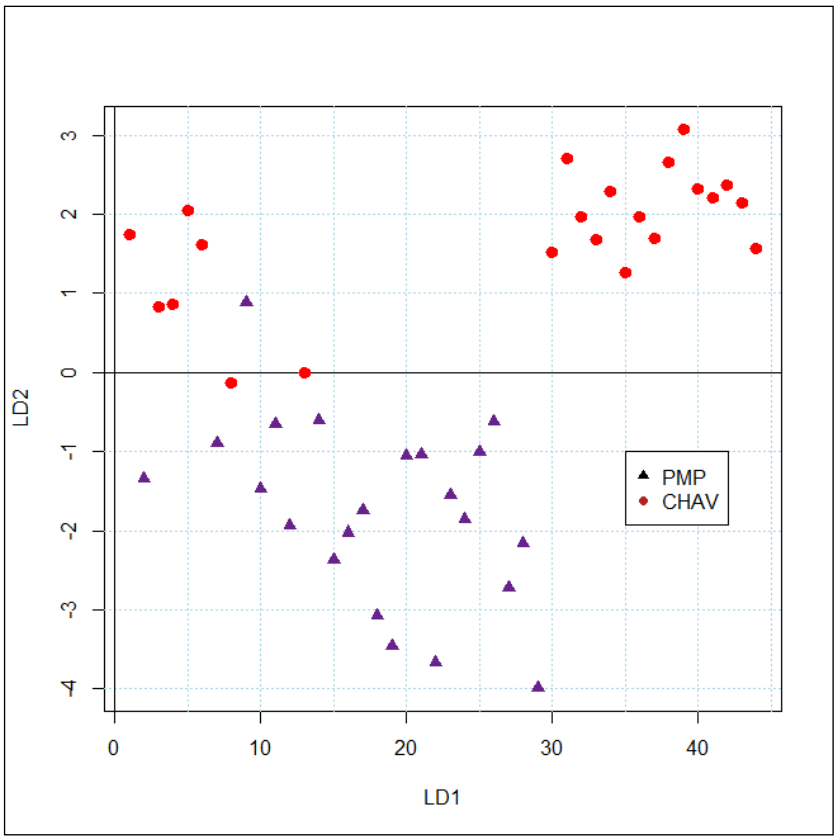


Fig. 5.37 Análisis discriminante de composición mineral de *Andropogon lateralis* en dos series de suelo. **PMP**: Pampín **CHAV**: Chavarría

En una segunda etapa se presentan los resultados de interés de la clasificación con análisis discriminante lineal para evaluar el desempeño del método en la matriz de confusión (Tabla 5.8).

Tabla 5.8 Matriz de confusión de Análisis Discriminante Lineal para *Andropogon lateralis*

	Chavarría	Pampín
Chavarría	75%	25%
Pampín	29%	71%
Exactitud: 73%		
Especificidad (TVN): 71%		
Sensibilidad (TVP): 75% Precisión: 60% AUC: 82%		

Para el entrenamiento de datos se utilizaron 32 muestras y para la prueba un total de 11 muestras. En los valores absolutos de la matriz de confusión se observa un error de Chavarría predicho como Pampín, y dos de Pampín predichas como Chavarría. Esto da como resultado un total de 3 muestras mal clasificadas, de un total de 11 muestras. Estos errores son consecuentes con los valores de la matriz de confusión presentada en la Tabla 5.8.

Los bajos valores que se observan en la matriz de confusión traen como consecuencia que las curvas ROC y PR estén lejos de ser los ideales para un clasificador binario. En las Fig. 5.38 y 5.39 se presentan la curva ROC y la curva PR correspondientes.

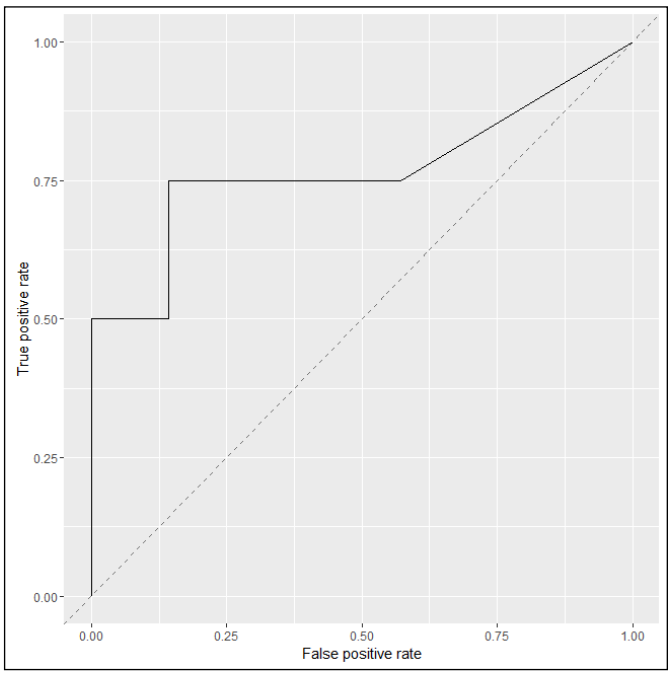


Fig. 5.38 Curva ROC correspondiente a *Andropogon lateralis* con análisis discriminante lineal

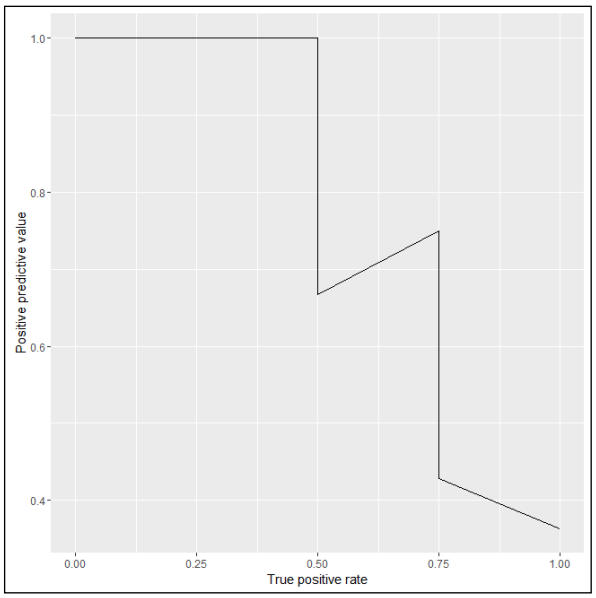


Fig. 5.39 Curva PR correspondiente a *Andropogon lateralis* con análisis discriminante lineal

Estos valores obtenidos para una clasificación binaria se consideran pobres por lo que a continuación se continuó en el análisis con otro algoritmo de clasificación para ver la posibilidad de un incremento del desempeño de los algoritmos de clasificación. Un algoritmo disponible que puede lidiar con el comportamiento no-lineal de las muestras es Support Vector Machines.

5.10.3.2 Support Vector Machines

En una siguiente etapa se probó un algoritmo de clasificación no lineal, para observar los porcentajes de clasificación y las métricas de interés. En una primera etapa se realizó una optimización de hiper-parámetros correspondientes al modelo. La optimización se realizó mediante búsqueda aleatoria (random search).

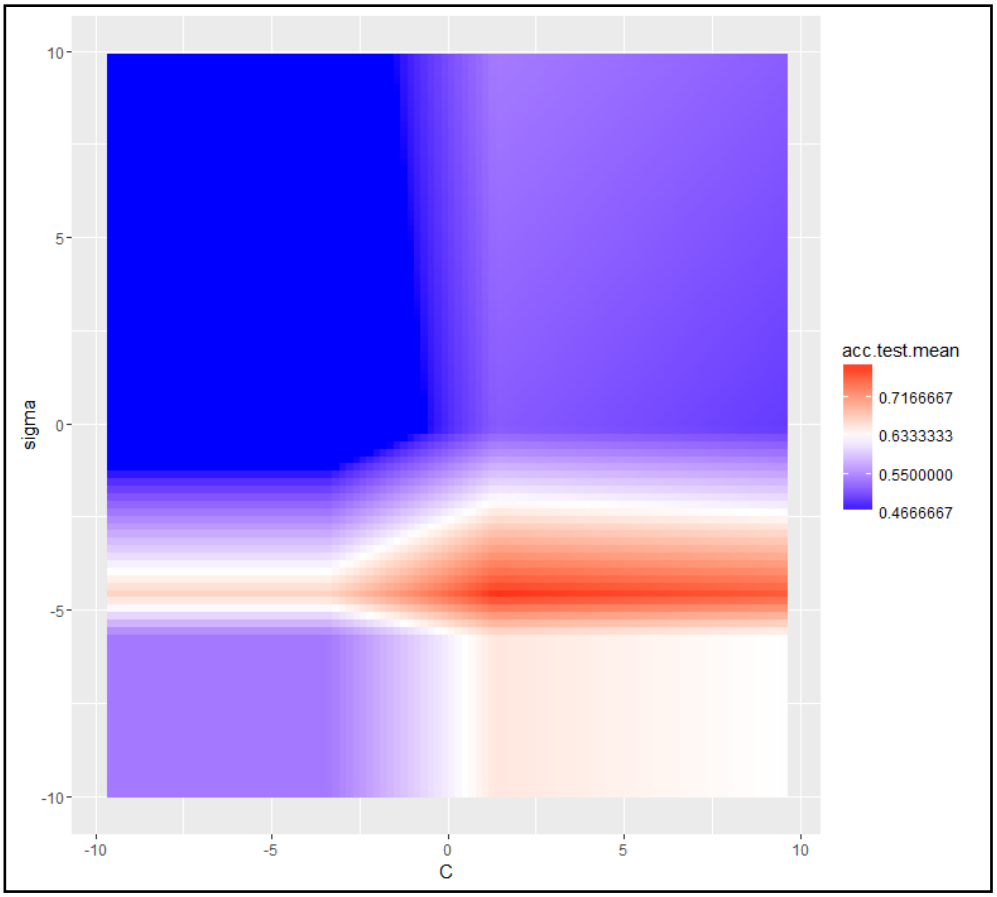


Fig. 5.40 Mapeo de posibles valores de C y sigma con su correspondiente para datos de *Andropogon lateralis*

Una vez ya con C y σ definidos (C: 6,38 y σ: 0,0835), vamos a evaluar el desempeño en el conjunto de datos de prueba.

A continuación en la Tabla 5.9, se presenta los valores de la matriz de confusión para evaluar el desempeño del algoritmo de clasificación.

Tabla 5.9 Matriz de confusión de Support Vector Machines para *Andropogon lateralis*

	Chavarría	Pampín
Chavarría	83%	17%
Pampín	20%	80%
Exactitud: 82%		
Especificidad (TVN): 80%		
Sensibilidad (TVP): 83% Precisión: 83% AUC: 86.6%		

Para el entrenamiento del modelo se utilizaron 32 muestras y 11 muestras para la prueba. A partir de la matriz de confusión confeccionada en la Tabla 5.9, se observa que de las 11 muestras del testeo, el modelo cometió dos errores totales, una de Chavarría predicha como Pampín, y una de Pampín clasificada como Chavarría. Esto trae como consecuencia valores de especificidad y sensibilidad muy similares entre sí, como así también, un valor de exactitud en el mismo orden. En las Fig 5.41 y 5.42 se presentan las curvas ROC y curva PR correspondientes al modelo SVM.

En la Fig. 5.41 se observa como a medida que el clasificador se acerca al 100% de área bajo la curva, ésta toma la forma de un clasificador ideal, al igual que la curva PR (Fig. 5.42).

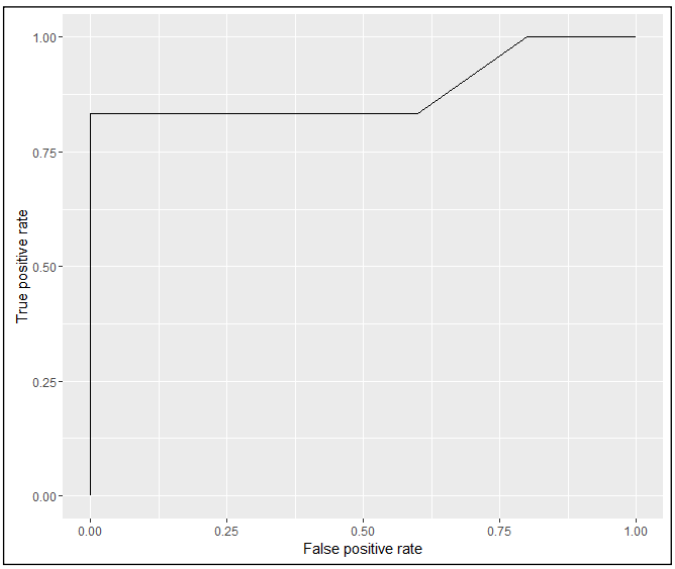


Fig. 5.41 Curva ROC correspondiente a *Andropogon lateralis* con support vector machines

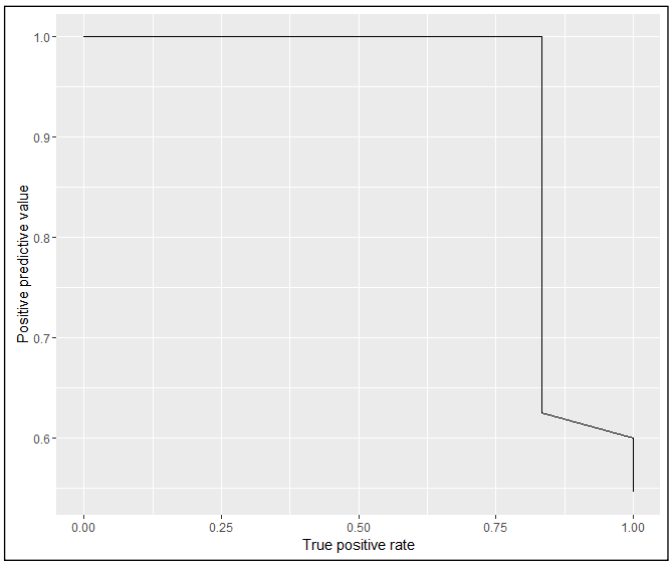


Fig. 5.42 Curva PR correspondiente a *Andropogon lateralis* con support vector machines

5.10.3.3 Random Forest

Random Forest es una técnica de ensamble que utiliza árboles de decisión para crear un meta-clasificador, que tiene mejor performance que los clasificadores solos por su cuenta (Raschka, S. 2015). Existen dos etapas importantes en el procedimiento del algoritmo de Random Forest que puede ser esquematizada en la Fig. 5.43

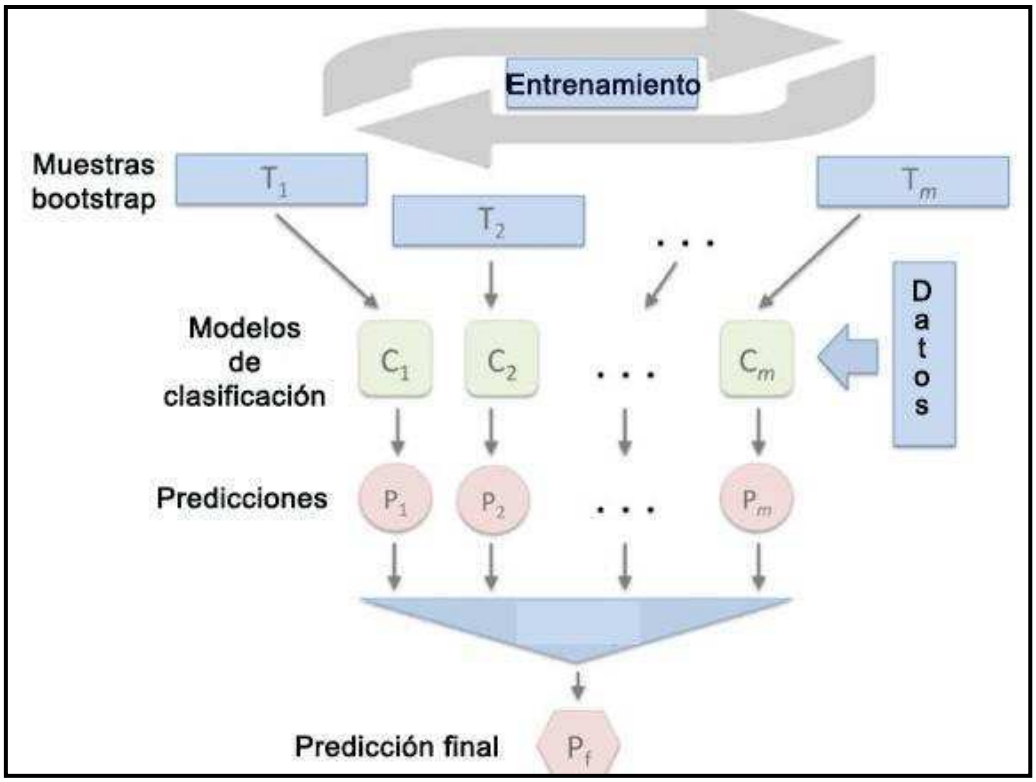


Fig. 5.43 Esquema de los pasos del algoritmo Random Forest

En una primera etapa existe la creación de nuevos set de datos a partir del conjunto de datos de entrenamiento. Esta etapa se denomina bootstrap aggregating, bootstrap resampling o simplemente bootstrap, ya que es un remuestreo con reemplazo (Forte, R. M. 2015, Raschka, S. 2015).

Luego se entrena el modelo correspondiente con los set de datos creados, y se anotan los resultados de las clases (en este caso, la serie de suelo) asignado por el modelo. Este proceso se repite *M* veces, para entrenar *M* modelos. En una etapa siguiente para cada observación del set de datos a través de los diferentes modelos, se computa la clase final a través del voto mayoritario. Para finalizar, se calcula la exactitud según el rótulo de la clase correspondiente y viendo la concordancia con la predicción.

A continuación se presentan los resultados de la matriz de confusión para la clasificación con Random Forests (Tabla 5.10).

Tabla 5.10 Matriz de confusión de Random Forest para *Andropogon lateralis*

	Chavarría	Pampín
Chavarría	80%	20%
Pampín	0%	100%
Exactitud: 91%		
Especificidad (TVN): 100%		
Sensibilidad (TVP): 80%		
Precisión: 100%		
AUC: 97%		

Para el entrenamiento del modelo se utilizaron 32 muestras y 11 muestras para la prueba. En cuanto a los errores en los valores absolutos de la matriz de confusión se observan un solo error en las muestras de Chavarría clasificada como Pampín. Esto trae como consecuencia que los valores porcentuales de la matriz de confusión en la tabla 5.10 sean altos en la gran mayoría de las métricas consideradas.

Los valores obtenidos con Random Forest fueron superiores a los obtenidos con SVM, en casi todas las métricas consideradas, a excepción de la sensibilidad. Esto trae como consecuencia que se considere a Random Forest como mejor clasificador que SVM.

Se observa que las curvas ROC y PR son consistentes con estos valores (Fig. 5.44 y 5.45), presentando los valores más altos a lo largo de todo el proceso de clasificación de muestras de *Andropogon lateralis.*

A continuación se observa la curva ROC correspondiente a un valor de 0,97.

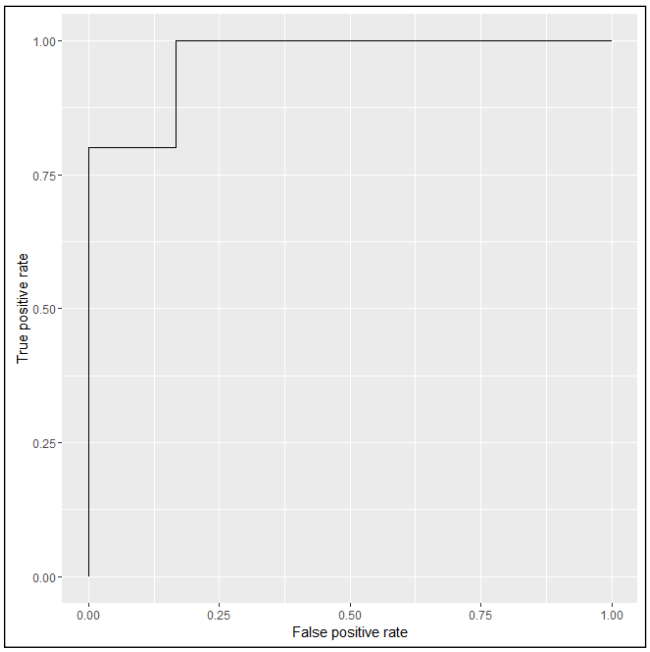


Fig. 5.44 Curva ROC correspondiente a *Andropogon lateralis* con Random Forest

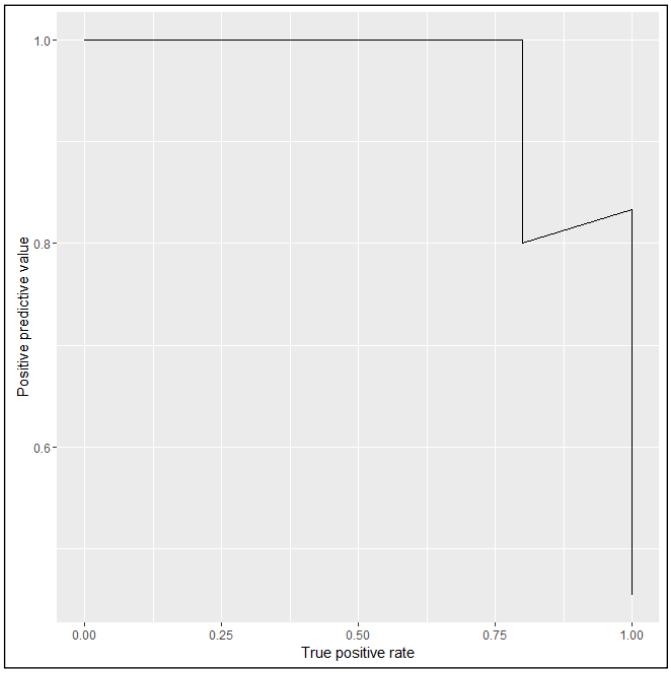


Fig. 5.45 Curva PR correspondiente a *Andropogon lateralis* con Random Forest

5.11 Clasificación según origen geográfico

Finalmente, se propone modelar los datos experimentales de acuerdo al origen geográfico de las muestras de partes aéreas correspondientes a cada especie vegetal estudiada. En esta etapa se aplicaron métodos clasificatorios del tipo multi-clase (más de dos clases o grupos). Con el objeto de evaluar el desempeño de cada algoritmo se

consideraron como métricas de interés: la exactitud global, la sensibilidad (SENS), la especificidad (SPEC), y el parámetro Kappa.

5.11.1 Desmodium incanum

Los resultados obtenidos al aplicar tres técnicas de análisis de datos: LDA, RF y SVM a los resultados de composición multielemental de las muestras de *Desmodium incanum* provenientes de los distintos sitios geográficos se resumen en la tabla 5.11. Para calcular estos resultados, se dividió el conjunto de datos, en datos de entrenamiento, y conjunto de datos de prueba, mediante la metodología de validación cruzada. Los sitios de muestreo fueron: PF (Paso Florentin), PN (Paso Naranjito), RP (Ramada Paso), SC (San Cosme) y SM (San Miguel).

Tabla 5.11 Resultados la clasificación de *Desmodium incanum* según origen geográfico

	LDA		SVM C=100	σ = 0.01	RF ntree=1000	mtry=1
	Sens	Espec	Sens	Espec	Sens	Espec
PF	100%	100%	100%	100%	100%	75%
PN	-	88%	100%	100%	-	100%
RP	100%	87.5%	100%	100%	100%	75%
SC	100%		100%	100%	-	100%
SM	33%	100%	100%	100%	33%	85%
Exactitud Global	70%		100%		50%	
Kappa	0.625		1		0.36	

En la tabla se expresan los valores de exactitud global, kappa, como así también la sensibilidad (Sens) y la especificidad (Espec) por cada grupo de muestras estudiado. También se observan los valores correspondientes al intervalo de 95% de confianza, y al igual que el estadístico Kappa. Tanto en el caso de SVM como RF se especifican los

valores de optimización del algoritmo; estos fueron para SVM (C = 100 y σ = 0.01) y para RF (ntree o número de árboles = 100 y mtry = 1). El estadístico Kappa toma vital importancia en los casos de clasificación en que las clases o grupos están desbalanceados, ya que en estos casos la exactitud podría ser alta simplemente clasificando bien las clases mayoritarias. Los valores de Kappa más cercanos a uno indican un perfecto acuerdo entre la predicción del modelo y los valores verdaderos (Lantz, B. 2015). Se observa que los valores más altos fueron obtenidos con SVM, en estos casos, siempre se elige el algoritmo más simple, esto se expresa claramente en la regla de Occam. La regla de Occam lo que dice es que el modelo más sencillo que se ajusta a los datos es también el más plausible (Abu-Mostafa, Y. S., et al. 2012).

5.11.2 Schizachyrium microstachyum

Los resultados obtenidos al aplicar tres técnicas de análisis de datos: LDA, RF y SVM a los resultados de composición multielemental de las muestras de *Schizachyrium microstachyum* provenientes de los distintos sitios geográficos se resumen en la tabla 5.12. Para calcular estos resultados, se dividió el conjunto de datos en datos de entrenamiento y set de datos de prueba, mediante la metodología de validación cruzada. Los sitios de muestreo fueron: PF (Paso Florentin), PN (Paso Naranjito), RP (Ramada Paso), SC (San Cosme) y SM (San Miguel).

Tabla 5.12 Resultados la clasificación de *Schizachyrium microstachyum* según origen geográfico

	LDA		SVM		RF	
			C=10	σ = 0.01	ntree=1000	mtry=2
	Sens	Espec	Sens	Espec	Sens	Espec
PF	100%	100%	100%	100%	100%	75%
PN	100%	100%	100%	100%	100%	100%
RP	-	100%	100%	100%	100%	75%
SC	100%	70%	100%	100%	100%	100%

SM	33%	100%	100%	100%	100%	100%
Exactitud Global		75%		100%		100%
Kappa		0.67		1		1

En la tabla se observan los valores correspondientes a la clasificación geográfica de muestras de *Schizachyrium microstachyum.* En la Tabla 5.12 se exponen los valores de exactitud global, kappa, sensibilidad y especificidad por grupo o por clases.

Se observa que los valores de exactitud obtenidos para los datos de prueba son iguales (100%) para SVM y RF, y menor (75%) para LDA. En este caso, al presentar dos algoritmos, un mismo valor de exactitud en la clasificación, a la hora de optar por uno se debe tener en cuenta, el tamaño del conjunto de datos, y el coste computacional que puede implicar el procesamiento de los datos. En general, el algoritmo SVM requiere mayor coste computacional y es recomendable para un conjunto de datos menor a 100.

5.11.3 Andropogon lateralis

En la tabla 5.13 se resumen los resultados correspondientes a *Andropogon lateralis* según origen geográfico, según tres técnicas de análisis de datos: LDA, SVM y RF. Para calcular estos resultados, se dividió el conjunto de datos en datos de entrenamiento y conjunto de datos de prueba, mediante la metodología de validación cruzada. Los sitios de muestreo fueron: PF (Paso Florentin), PN (Paso Naranjito), RP (Ramada Paso), SC (San Cosme) y SM (San Miguel).

Tabla 5.13 Resultados la clasificación de *Andropogon lateralis* según origen geográfico

	LDA	SVM		RF	
		C=10	σ = 0.01	ntree=1000	mtry=2

	Sens	Espec	Sens	Espec	Sens	Espec
PF	100%	100%	100%	100%	100%	75%
PN	75%	100%	100%	100%	100%	100%
RP	100%	100%	100%	100%	100%	75%
SC	100%	88%	100%	100%	100%	100%
SM	100%	100%	100%	100%	100%	100%
Exactitud Global	90%		100%		100%	
Kappa	0.87		1		1	

Se observan que los valores óptimos fueron los obtenidos con RF y SVM. Como se mencionó anteriormente, a la hora de optar por un algoritmo de clasificación cuando presentan los mismos valores de exactitud, es conveniente tener en cuenta el tipo de dato (tensores, matrices) con el que se está trabajando y el coste computacional que eso implicaría. Los valores de Kappa son menores que la exactitud en el caso de LDA, pero nos dan una real dimensión del porcentaje de exactitud en la clasificación al tener clases desbalanceadas.

5.12 Referencias bibliográficas

Abu-Mostafa YS, Magdon-Ismail M, Lin H-T. (2012) Learning from Data: A Short Course: AMLBook.com.

Balbuena O ML, Lucian C, Conrad J, Wilkinson N, Martin F. (2013) Estudios de la Nutrición Mineral de los Bovinos para carne del este en las provincias de Chaco y Formsosa (Argentina). Rev Veterinaria Argentina 56: 25-32.

Bernardis A, Villafañe R, Pellerano R, Marchevsky E. (2016) Perfil mineral en los pastizales de Andropogon lateralis y Sorghastrum setosum (Gramineae) en Corrientes, Argentina. Revista Facultad de Ciencias Agrarias UNCuyo.

Brereton RG. (2009) Chemometrics for Pattern Recognition: John Wiley & Sons.

Forte RM. (2015) Mastering Predictive Analytics with R: Packt Publishing Ltd.

Goodfellow I, Bengio Y, Courville A. (2016) Deep Learning: MIT Press.

Gupta UC, Wu K, Liang S. (2008) Micronutrients in Soils, Crops, and Livestock. Earth Science Frontiers.15:110-125.

Hackeling G. (2014) Mastering Machine Learning with scikit-learn: Packt Publishing Ltd.

Japkowicz N, Shah M. (2011) Evaluating Learning Algorithms: A Classification Perspective: Cambridge University Press.

Kirk M. (2014) Thoughtful Machine Learning: A Test-Driven Approach: O'Reilly Media, Inc.

Kruzlicova D, Fiket Ž, Kniewald G. (2013) Classification of Croatian wine varieties using multivariate analysis of data obtained by high resolution ICP-MS analysis. Food Res. Int.54:621-626.

Kumar N, Bansal A, Sarma GS, Rawal RK. (2014) Chemometrics tools used in analytical chemistry: An overview. Talanta.123:186-199.

Lantz B. (2015) Machine Learning with R: Packt Publishing Ltd.

Mufarrege D. (1999) Los minerales en la alimentación de vacunos para carne en la Argentina. EEA INTA Mercedes, Corrientes, Trabajo de Divulgación Técnica.

Mufarrege D. (2000) El contenido de zinc de las pasturas naturales en la provincia de Corrientes y en la región del NEA. EEA INTA Mercedes, Corrientes, Noticias y Comentarios N° 341.

Mufarrege D. (2003) El cobre en la ganadería del NEA. INTA EEA Mercedes, Corrientes Noticias y Comentarios N° 381

Moscuzza CH, Pérez-Carrera AL, Volpedo AV, Fernández-Cirelli A. (2012) Forage enrichment with copper and zinc in beef grazing systems in Argentina. J Geochem. Explor.121:25-29.

Peng RD. (2012) Exploratory Data Analysis with R: LeanPub

Raschka S. (2015) Python Machine Learning: Packt Publishing Ltd.

Varmuza K, Filzmoser P. (2009) Introduction to Multivariate Statistical Analysis in Chemometrics: CRC Press.

Watson CA, Öborn I, Edwards AC, Dahlin AS, Eriksson J, Lindström BEM, Linse L, Owens K, Topp CFE, Walker RL. (2012) Using soil and plant properties and farm management practices to improve the micronutrient composition of food and feed. J Geochem. Explor.121:15-24.

Zheng, A. (2015) Evaluating Machine Learning Models USA: O'Reilly Media, Inc.

SECCIÓN IV: COMENTARIOS FINALES

6) Capítulo VI

La Espectrometría de Masas por Plasma Acoplado Inductivamente (ICP-MS) permitió cuantificar las concentraciones de 18 elementos (Al, B, Cd, Co, Cr, Cu, Li, Mo, Ni, Rb, Sb, Se, Sn, Sr, Ti, Tl, V y Zn) en las partes aéreas de 42 muestras de *Desmodium incanum*, 48 muestras de *Schizachyrium microstachyum* y 43 muestras de *Andropogon lateralis*. La técnica analítica propuesta demostró tener rangos de exactitud y sensibilidad adecuados para la determinación de concentraciones para las especies en estudio.

Para la mejor comprensión de los valores se realizó un análisis exploratorio de datos que incluyó la presentación de los datos de forma gráfica, usando gráficos de Cajas y Bigotes que permiten observar la distribución de valores en las muestras estudiadas. En la siguiente etapa se realizó un análisis de componentes principales (PCA) para permitir una mejor comprensión de las posibles relaciones entre variables, y descartar la presencia de datos extremos. Los resultados obtenidos por la PCA, fueron corroborados por análisis de conglomerados (HCA) de las variables estudiadas.

Los resultados de PCA de *Desmodium incanum* permitió una reducción de variables de la PC1: 29,2% y de la PC2: 12, 6%, en los gráficos de loadings y de sedimentación. El gráfico de scores nos permite visualizar la distribución de algunas muestras de Chavarría entre las de Pampín. Luego el análisis HCA permitió agrupar las variables en concordancia con lo presentado en el gráfico de loadings de la PCA, en cuatro grupos. Estos grupos fueron; Grupo 1: Al, Sr, B, V, Rb, Zn y Co; Grupo 2: Cd y Mo; Grupo 3: Cu, Se, Li, Sn, Ti y Sb; Grupo 4: Ni, Tl y Cr.

La aplicación de PCA a muestras de *Schizachyrium microstachyum* permitió una reducción a dos de las primeras componentes con valores de PC1: 23% y de PC2: 10, 1%. El gráfico de scores permitió una visualización de uno o más grupos más o menos uniformemente distribuido en muestras de Pampín, cosa que no sucede con las muestras de Chavarría. Los resultados de HCA de *Schizachyrium microstachyum*, permitieron agrupar las variables en estudio en cuatro grupos. Éstos fueron; Grupo 1: Al, Cu, Se, Sn, Li, Tl, Mo y Cr; Grupo 2: B, Rb y Ti; Grupo 3: Zn y Sb; Grupo 4: Cd, Sr, V, Co y Ni. Estos resultados fueron consistentes con los presentados en el gráfico de loadings de PCA con algunas diferencias.

Al aplicar el análisis de componentes principal a muestras de *Andropogon lateralis* permitió una reducción de dimensiones PC1 y PC2 con valores de 21,7% y 10,5% respectivamente. La distribución de muestras en el gráfico de scores permitió visualización en el espacio dimensional PC1-PC2 observándose dos grupos, pero algunas muestras Pampín están próximas a las de Chavarría. Luego de aplicar HCA a las variables en estudio, se agruparon las variables en cuatro grupos. Estos grupos fueron; Grupo 1: Al, Cu, Se, Sn, Li y Ti; Grupo 2: Tl, Mo y Cr; Grupo 3: B, Rb, Zn y Sb; Grupo 4: Cd, Sr, V, Co y Ni.

Una vez completado el análisis exploratorio de datos, se propusieron modelos clasificatorios usando distintos criterios para agrupar a las muestras, ya sea de acuerdo a la serie de suelo donde fueron recolectados las plantas o el origen geográfico de las mismas. Se utilizaron tres diferentes tipos de algoritmos con características diferentes de acuerdo al grado de complejidad para la clasificación de los datos en distintos grupos, los métodos utilizados ordenados secuencialmente de acuerdo a su

complejidad fueron: LDA (análisis discriminante lineal), SVM (support vector machines) y RF (random forest).

Según la distribución de puntos observada en los análisis exploratorios, se hizo necesaria el empleo de algoritmos lineales más simples como LDA o más complejos que no poseen supuestos de linealidad a la hora de encontrar una función que pueda discriminar según serie de suelo.

Los resultados obtenidos al clasificar las partes aéreas de *Desmodium incanum* con análisis discriminante lineal fueron una exactitud total de 90%, una sensibilidad de 100% y un área bajo la curva de 90%. Estos valores nos indican que la clase considerada como positiva (en este caso Chavarría), mostraron una clasificación 100% correcta de las muestras de Chavarría, y 86% para las muestras de Pampín. Estos resultados no hicieron necesario el empleo de otro algoritmo más complejo en búsqueda de mejores resultados.

Los resultados obtenidos de la clasificación de *Schizachyrium microstachyum* con análisis discriminante lineal fueron una exactitud de 67%, una sensibilidad de 83% y un área bajo la curva de 63%. Estos valores nos indican una clasificación de 67% para la clase de Chavarría, considerada como positiva, que se refleja en el valor de sensibilidad. Este valor disminuye finalmente con respecto la exactitud ya que la exactitud considera también las muestras de Pampín, que fueron correctamente clasificadas en un 67%. Estos valores son buenos pero no son óptimos para una clasificación binaria. Luego, se probó otro algoritmo de aprendizaje llamado SVM. Los resultados obtenidos fueron una exactitud de 92%, una sensibilidad de 83% y un área bajo la curva de 91,6%. Estos resultados nos indican que la clase considerada como

negativa (Pampín), tuvo un porcentaje de muestras clasificadas correctamente mayor que las de Chavarría lo que se expresa en un valor de exactitud mayor que la sensibilidad.

Los resultados obtenidos al aplicar análisis discriminante lineal a muestras de *Andropogon lateralis* fueron una exactitud de 73%, una sensibilidad de 75% y un área bajo la curva de 82%. En búsqueda de mejores valores de exactitud se aplicó el algoritmo SVM para lograr mejores valores de clasificación. Los resultados obtenidos fueron una exactitud de 82%, una sensibilidad de 83% y un área bajo de la curva de 86,6%. Estos valores si bien son buenos no son óptimos para un clasificador binario, por ello se recurrió a otro método de clasificación que trabaje con límites de decisión no lineales. En una tercera instancia, los datos fueron sometidos a una clasificación con RF, obteniéndose un valor de exactitud de 91%, una sensibilidad de 80% y un área bajo la curva de 96,6%. Un valor de exactitud mayor se observa debido a un valor de 100% en las muestras de especificidad, en este caso, las muestras de Pampín.

Estos valores de exactitud fueron más altos que los algoritmos previos, lo que permite afirmar que Random Forest fue el mejor clasificador para el conjunto de datos de *Andropogon lateralis*.

Finalmente, se estudió el comportamiento de estas mismas especies forrajeras, esta vez teniendo como criterio de clasificación el sitio geográfico. Ya en esta etapa se observó el desempeño de estos tres algoritmos mencionados, arrojando mejores valores de clasificación, aquellos que tienen como límite de decisión una función no lineal (SVM y RF) en la mayoría de los casos.

Para las muestras de *Desmodium incanum* los valores de clasificación para el conjunto de prueba de LDA fueron 70% de exactitud y kappa de 0.625. Los valores de clasificación de SVM fue de 100% pero los valores de RF resultaron muy bajos (50%).

Para las muestras de *Schizachyrium microstachyum* los mejores valores de clasificación fueron obtenidos mediante SVM y RF, con valores de exactitud de 100% y de Kappa de 1. En estos casos se elige el algoritmo de clasificación más simple o el algoritmo que menos coste computacional implique.

Para las muestras de *Andropogon lateralis* los mejores valores de exactitud y Kappa fueron para el algoritmo SVM y RF con valores de exactitud de 100% y de Kappa de 1 lo que es consistente con los análisis previos según serie de suelo.

Brevemente, a modo de conclusión, se puede decir que al aplicar clasificación binaria de las muestras, se aplicaron algoritmos de complejidad creciente para obtener un valor aceptable en la discriminación de las muestras (LDA para muestras de Desmodium incanum, SVM para Schizachyrium microstachyum y RF para Andropogon lateralis).

En cuanto a la clasificación multiclase de las muestras, el mismo algoritmo (SVM) se utilizó para lograr una clasificación óptima de todas las muestras, ya que RF tiene leves errores en los valores de especificidad en las tablas de confusión.

En base a los resultados obtenidos en este trabajo, se puede concluir que los mayores aportes de este trabajo radican en la profundización del conocimiento de la composición mineral de especies forrajeras de elementos de interés nutricional como también elementos a nivel de vestigios sin función fisiológica demostrada o conocida. También es importante destacar la propuesta de modelos quimiométricos que

permiten establecer con seguridad la procedencia geográfica de estas especies en estudio. Esta metodología estadística puede aplicarse a otros productos vegetales, no solo para clasificación geográfica, sino dentro de un sistema de trazabilidad de productos regionales en vistas de su eventual exportación.