VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Revisión exploratoria de literatura científica en acuicultura: Análisis de tendencias utilizando un modelo probabilístico bayesiano y herramientas de machine learning

Autor: Javier Antonio De La Hoz Maestre
Tutora: Dra Maria José Fernández Gómez

2020

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

Dpto. de Estadística
Universidad de Salamanca

DRA MARÍA JOSÉ FERNÁNDEZ GÓMEZ

Profesora titular del Departamento de Estadística de la Universidad de Salamanca

---

CERTIFICA que **D./D.ª Javier Antonio De La Hoz Maestre** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo que para optar título de Máster en Análisis Avanzado de Datos Multivariantes y Big Data presenta con el título ***"Revisión exploratoria de literatura científica en acuicultura: Análisis de tendencias utilizando un modelo probabilístico bayesiano y herramientas de machine learning"***, autorizando expresamente su lectura y defensa.

Y para que conste, firma el presente certificado en Salamanca a 13 de julio de 2020.

María José Fernández Gómez

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Revisión exploratoria de literatura científica en acuicultura: Análisis de tendencias utilizando un modelo probabilístico bayesiano y herramientas de machine learning

Trabajo para optar al título de Máster en
Análisis Avanzado de Datos Multivariantes y Big Data
por la Universidad de Salamanca.

Presenta:

Javier Antonio De La Hoz Maestre

Salamanca

2020

DEDICATORIA

```
1 dedicate <- function(reader){
2    if (reader == mywife){
3        print("To my wife...")
4    }
5    else if (reader == myparents){
6        print("To my parents...")
7    }
8    else if (reader == myteachers){
9        print("To my teachers...")
10   }
11   else{
12       print("To all my friends...")
13   }
14   }
```

ABSTRACT

Research in aquaculture develops fast as it has to respond to the significant growth of this industry; there are many biological challenges, technical improvements and technology developments that need to be addressed by researchers. Thousands of articles on aquaculture have been published, so it is laborious and time consuming to extract information from accumulated collections. The aim of this study was to understand the distribution patterns and trends of the literature available in the field of aquaculture in order to improve knowledge, nature and structure of these publications. This study performed a literature review of 38319 abstracts published in 14 top-tier aquaculture journals, between the years 1972 and 2019. A Latent Dirichlet Allocation (LDA) was applied to perform text mining on the dataset, finding 40 key topics. Machine learning tools were used in the subsequent distribution and composition of words. As result, we found that topic modeling has the ability to segregate a collection of articles on different topics, and could be used as a tool to understand literature, not only recapturing known facts but also discovering other relevant topics. In general, the topics found confirm key areas of aquaculture research that have been identified by qualitative studies. However in our case it also provides a quantitative evaluation and analysis in the most recent scientific literature.

Keywords: Topic model, Text mining, Aquaculture Research, Latent Dirichlet Allocation, Published articles.

TABLA DE CONTENIDO

pág.

LISTA DE FIGURAS

LISTA DE TABLAS

1 INTRODUCCIÓN

La revisión de la literatura se considera parte integral dentro del proceso de investigación en cualquier área científica. Vom Brocke *et al*. (2009) afirman que la razón por la cual la revision de literatura ha jugado un papel decisivo es porque la ciencia sigue siendo, ante todo, un esfuerzo acumulativo, ya que se crean nuevos conocimientos a partir del proceso de interpretación y combinación de conocimientos existentes.

Paré *et al*. (2015) resumen el propósito de una revision en las siguientes categorías que reflejan las razones por la cuál se realiza una revisión concreta de la literatura ~~dada~~:
(i) identificar lo que se ha escrito sobre tema, (ii) determinar en qué medida un área de investigación específica revela tendencias o patrones interpretables, (iii) agregar hallazgos empíricos relacionados con una pregunta de investigación para apoyar la práctica basada en evidencia, (iv) generar nuevos marcos y teorías e (v) identificar temas o preguntas que requieren más investigación.

Por lo general, se utilizan dos tipos principales de búsquedas bibliográficas sistemáticas para realizar una revisión de la literatura: búsquedas de palabras clave y búsquedas de citas hacia adelante/hacia atrás (Adams J, *et al.,* 2007).

Una búsqueda de palabras clave comienza con la identificación de una lista de términos de búsqueda que se cree que representan adecuadamente las palabras que los autores

pueden usar para abordar un tema de interés, luego con el uso de operadores de lógica booleana, en una bases de datos indexada, como *Web of Science* o *Scopus*, se recupera el conjunto de artículos de interés para su posterior análisis. El procedimiento descrito se realiza en muchos casos de forma iterativa para garantizar una mejor coincidencia. El reconocimiento de las debilidades de esta forma de búsqueda no son recientes puesto que Garfield (1955) reconoce las limitaciones de las búsquedas con lógica booleana.

En el procedimiento de búsqueda de citas hacia adelante y hacia atrás, inicialmente se selecciona un número de artículos que se consideran esenciales en el campo de interés. Posteriormente se realizan búsquedas hacia adelante y hacia atrás, seleccionando artículos que citan a los artículos del conjunto de referencia inicial. Al no basarse en palabras clave que pueden ser compartidas por distintas áreas de la ciencia esta forma es ventajosa cuando se revisan temas latentes. Sin embargo, las búsquedas de citas hacia adelante y hacia atrás se basan en trabajos previos que ya han unido múltiples disciplinas (Wang *et al.*, 2011)

Ambos tipos de búsquedas tienen limitaciones cuando se revisan tópicos latentes abordados por múltiples campos científicos (de Wildt *et al.*, 2018). Aunado a lo anterior, ambos procedimientos implican trabajo considerable, pues leer individualmente una gran cantidad de documentos tendrá un alto costo de tiempo para el investigador, lo que conlleva a limitar la cantidad de documentos para revisar. Esta fase exploratoria es un problema dado que lo se necesita es una visión general de las direcciones de investigación.

La investigación científica global ha crecido significativamente en las últimas tres décadas (National Science Board, 2012). Este fuerte aumento también incluye la producción de documentos en el área de la acuicultura (Mather *et al.*, 2008; Natale *et al.*, 2012). En la actualidad, un número creciente de investigadores dedica sus esfuerzos a estudiar diferentes aspectos de la acuicultura para mejorar las técnicas y protocolos de cría de las especies cultivadas; además, existe la necesidad de diversificar la acuicultura, ya que solo unas pocas especies, especialmente marinas, pueden reproducirse y criarse masivamente en cautiverio (Natale *et al.*, 2012).

La investigación acuícola involucra áreas muy diversas (ingeniería, ecología, biología, fisiología, economía, ciencias ambientales y políticas, entre otras) que, en la mayoría de los casos, deben desarrollarse conjuntamente para producir con éxito una especie específica a nivel industrial. La investigación y, por lo tanto, el logro de resultados, en muchos casos, se realizan en diferentes partes del mundo, a una velocidad significativa. Además, los nuevos temas y tendencias comienzan a surgir diariamente, causando superposiciones de investigación, y con esto, la pérdida de fondos que en la mayoría de los países es muy difícil de obtener. Por lo tanto, el estudio de la dinámica de investigación en acuicultura es crucial para aunar esfuerzos, articular diferentes campos y establecer nuevas líneas de investigación para el desarrollo de nuevas técnicas y tecnologías en esta industria.

Los avances en el aprendizaje automático y el procesamiento del lenguaje natural nos proporcionan una serie de técnicas que permiten sustituir el uso del tiempo del investigador por el uso del tiempo de la computadora. Por ejemplo, el enfoque del modelo

de tópicos (Blei & Lafferty, 2009) tiene la capacidad de encontrar los tópicos latentes subyacentes, o grupos de temas relacionados. Esto indica un potencial adecuado para el uso de modelado de tópicos en revisiones exploratorias de literatura en cualquier campo científico incluída la acuicultura.

Este documento se ha organizado de la siguiente manera:

- A la introducción le sigue el marco teorico, en donde se explican las consideraciones teóricas en las que se sustenta la investigación, se presentan, resumen y se comparan los conceptos básicos de los modelados de tópicos: Análisis Semántico Latente (LSA), Análisis Semántico Latente probabilistico (PLSA) y Asignación de Dirichlet Latente (LDA);
- A continuación se presentan las metas específicas que se pretenden alcanzar con el establecimiento de los objetivos;
- En la cuarta parte se presentan, desde el punto de vista metodológico, los métodos estadísticos que se utilizarán para el cumplimiento de nuestros objetivos;
- La quinta parte presenta los resultados de la investigación;
- Se finaliza con la presntación de las conclusiones.

2 MARCO TEORICO

Feldman *et al.* (1995) describió la minería de textos como "El proceso de extraer patrones interesantes de colecciones de textos muy grandes para el propósito de descubrir conocimiento" (Berry & Kogan, 2010).

En el campo de la minería de textos, expertos en aprendizaje automático, han investigado y presentado modelos de tópicos utilizando algoritmos de aprendizaje automático sin supervisión con el fin de descubrir de forma automática la información oculta en el texto (Blei., D.M., 2012 ; Blei, D. M., & Lafferty, J. D. 2009 ; Chemudugunta, C. 2010).

El modelado de tópicos es un problema clásico en el procesamiento del lenguaje natural y el aprendizaje automático. Hace referencia a un conjunto de algoritmos y métodos estadísticos de aprendizaje, reconocimiento y extracción de la información?que tienen como objetivo analizar la estructura oculta de una colección de documentos para descubrir los tópicos, cómo éstos se conectan entre sí y cómo cambian con el tiempo. La principal ventaja es que no se requieren anotaciones previas o etiquetado de los documentos, sino que los tópicos surgen del análisis de los textos originales (Blei., D.M., 2012).

En el modelado de tópicos la idea es encontrar un marco de modelado adecuado de datos discretos, en donde uno de los objetivos, al igual que en muchas técnicas

multivariadas, es la reducción de la dimensionalidad de los datos. Como resultado, algunos investigadores han propuesto reducir cada documento a un vector de números reales, cada uno de los cuales representa proporciones de recuentos. Todos los modelos de tópicos están basados en las mismas suposiciones básicas, es decir, cada documento consiste en una mezcla de tópicos y cada tópico consiste en una colección de palabras. Los campos de aplicación cubren casi todas las áreas de minería de texto y procesamiento de la información, como puede ser el resumen de texto, la recuperación de información y la clasificación de textos, entre otras (Kao, A., & Poteet, S. R., 2007). El modelado de tópicos nos permite organizar y resumir archivos electrónicos en diversos formatos (páginas web, artículos científicos, libros, imágenes, sonido, videos y redes sociales) a una escala que sería imposible de llevar a cabo mediante ~~por~~ la anotación humana (Blei, 2012).

En la actualidad, dentro los algoritmos utilizados para el modelado de tópicos se destaca el Asignacion Latente de Dirichlet (LDA) propuesto por Blei *et al.* (2003) considerado como uno de los más populares (Jacobi, *et al*., 2016; Blei, 2012; DiMaggio *et al.,* 2013; Grimmer, 2010) , el Análisis Semántico Latente (LSA) (Deerwester et al.,1990) y el Análisis Semántico Latente Probabilístico (PLSA) (Hofmann, 1999). Estos modelos junto con frecuencia de término – frecuencia inversa de documento TF-IDF (Salton *et al.,* 1975) serán descritos a continuación.

En las siguientes secciones utilizaremos la siguiente nomenclatura y definiciones que son comunes entre los diversos algoritmos o técnicas de modelado de tópicos, en especial las definiciones de Blei *et al*., (2003):

- Una ***palabra*** es la unidad básica de datos discretos, definida como un elemento de un vocabulario indexado por {1, ..., V}. Representamos palabras usando vectores unitarios que tienen un solo componente igual a uno y el resto igual a cero. Se usarán superíndices para denotar componentes. La palabra v-esíma en el vocabulario está representada por un vector V tal que $w^v = 1$ y $w^u = 0$ para $u \neq v$;
- Un ***documento*** es una secuencia de N palabras denotadas por $w = (w_1, w_2, ..., w_N)$, donde w_n es la n-ésima palabra en la secuencia;
- Un ***corpus*** es un subconjunto de M documentos, construido de acuerdo con una serie de criterios de diseño explícitos para un propósito específico (Atkins, *et al.*, 1992) denotados por $D = \{w_1, w_2, ..., w_M\}$;
- Un ***tópico*** es una variable latente denotada por $Z = \{z_1, z_2, ... z_k\}$ donde el número k es un parámetro que debemos especificar.

2.1 FRECUENCIA DE TÉRMINO – FRECUENCIA INVERSA DE DOCUMENTO (TF-IDF)

TF-IDF (Salton *et al.,* 1975), es un modelo algebraico para representar tanto documentos de texto como vectores. Se basa en dos supuestos (Sebastiani, 2002). El primero es que cuanto más frecuente es un término en un documento, más representativo es el contenido de este documento, que puede medirse por la Frecuencia de Término (TF); es decir, TF es la cantidad de veces que un término aparece en un documento. La segunda es que, en cuántos más documentos aparecen un término,

menor es el efecto que tiene dicho término a la hora de discriminar la importancia del documento y se presenta por la Frecuencia de Documento Inversa (IDF). En otras palabras, IDF significa que, los términos que aparecen con mucha frecuencia en un documento disminuyen el peso de este término y los términos que ocurren raramente aumentan el peso del término correspondiente, IDF se define como :

$$idf_i = log \frac{|D|}{|\{d:t_i \in d\}|} \quad (1)$$

donde $|D|$ es el número total de documentos en el corpus y $|\{d: t_i \in d\}|$ es el número de documentos en los que aparece el término t_i. Si el termino t_i aparece en todos los documentos del corpus, idf_i es igual a 0. El valor idf_i será mayor si el término t_i aparece en pocos documentos.

En otras palabras es una medida del poder discriminante de un término con respecto a un documento en un corpus; podemos asumir el valor TF-IDF, de la siguiente manera:

$$t_d = f_{t,d} * llog \frac{|D|}{|\{d:t_i \in d\}|} \quad (2)$$

donde $f_{t,d}$ es igual al número de veces que el término *t* aparece en el documento *d*.

2.2 ANÁLISIS SEMÁNTICO LATENTE (LSA)

El Análisis Semántico Latente (LSA) fue introducido por Deerwester *et al.* (1990). La idea básica del algoritmo del LSA es que un fragmento textual puede ser representado como una ecuación lineal, cuyo significado correspondería a la suma de los significados de las palabras que lo conforman esto es, a la frecuencia con que ellas co-ocurren en ese fragmento (Landauer *et al.*, 1998). La ecuación lineal se resuelve a través de la descomposición en valores singulares (SVD), que genera como resultado un espacio multidimensional compuesto por la representación matemático-vectorial de palabras y

documentos en un espacio semántico latente (Landauer *et al.*, 1998). En otras palabras, LSA es un análisis de componentes principales aplicado a la matriz término-documento (dtm) que es una matriz de dimensión n (documentos) x m (palabras), donde cada elemento a_{ij} es usualmente definido por una frecuencia ponderada del término i en el documento j. LSA utiliza la SVD de dicha matriz para identificar un subespacio lineal de baja dimensión, en el espacio de características TF-IDF de forma que se capture la mayor parte de la variación en la colección de documentos.

El algoritmo para LSA consta de las siguientes etapas (Evangelopoulos *et al.*, 2012):

(i) Creación de la matriz de entrada. Se genera una matriz de términos del documento, en donde las celdas se utilizan para representar la importancia de las palabras en las oraciones. Existen diferentes enfoques para los valores de las celdas. Los más usados son:

- Frecuencia de la palabra: la celda se completa con la frecuencia de la palabra en la oración;
- Representación binaria: la celda se rellena con 0/1 dependiendo de la aparición (o no) de una palabra en la oración;
- TF-IDF: la celda se rellena con el valor TF-IDF de la palabra. Un valor TF-IDF más alto significa que la palabra es más frecuente en la oración, pero menos frecuente en todo el documento. Cuanto más alto es el valor, más representativa es la palabra para esa oración;

- Log entropía (Shannon, 1948): la celda se completa con el valor de log-entropía de la palabra, que proporciona información sobre cómo de informativa es la palabra en la oración.

Las matrices generadas a través de este paso tienden a ser de gran dimensionalidad, y con un alto porcentaje de los valores de las celdas iguales a cero, de ahí el apelativo *matriz sparse* (matrices dispersas). En este punto, destacamos que LSA elimina las palabras usadas más frecuentes, conocidas como *stopword* que habitualmente son artículos y preposiciones, es decir palabras que no aportan significado.

(ii) Reducción de la dimensionalidad: en este paso se realiza la SVD en la matriz términos-documento (dtm) generado. La idea básica es que los elementos de la matriz de términos del documento se pueden representar como puntos en el espacio euclidiano, en donde los vectores se utilizan para mostrar los documentos u oraciones en este espacio. La SVD descompone la matriz dtm X_{txd}, en el producto de otras tres matrices: una matriz U_{txm} columna-ortogonal con m que representa la dimensionalidad, una matriz diagonal S_{mxm} con valores singulares dispuestos en orden decreciente, y, una matriz transpuesta V_{dxm} columna-ortogonal (Lochbaum y Streeter, 1989), donde t denota el número de términos y d denota el número de documentos. Posteriormente, las matrices se truncan en un número k arbitrario de dimensiones, para eliminar parte del ruido existente en la matriz original y así extraer la relación semántica latente en la colección (Figura 1).

(iii) Análisis cuantitativo: finalmente, los términos y documentos representados en el espacio definido por las k dimensiones retenidas mediante la SVD se analizan mediante la implementación de un método analítico específico en función del objetivo buscado. Por ejemplo, si el objetivo es la comparación de documentos, la evaluación de documentos, su clasificación o la coherencia entre documentos, utilizaremos como medida de similaridad el coseno del ángulo entre los vectores fila que representan a los documentos, mientras que si el objetivo es la categorización de documentos o resumen de los mismos, utilizaremos *un análisis de clusters* o un análisis factorial.

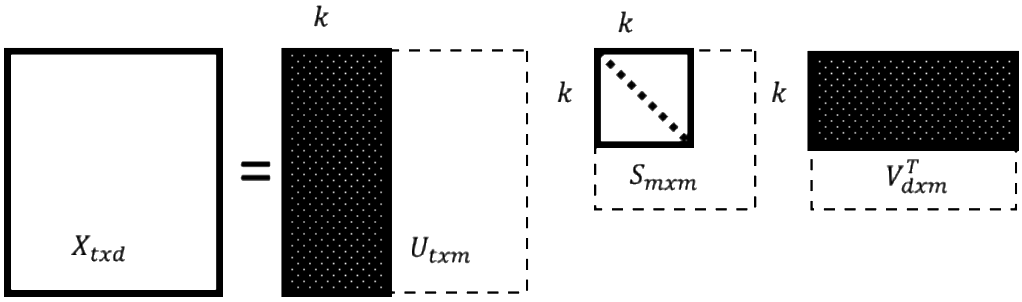

Figura 1. Descomposición y truncamiento de la matriz términos-documento (adoptado de Berry *et al.*, 1995)

2.3 ANÁLISIS SEMÁNTICO LATENTE PROBABILÍSTICO (PLSA)

El modelo anterior se basó en un enfoque determinista, pero también es posible definir un modelo probabilístico sobre el espacio determinado por documentos y palabras. Dicho modelo es el denominado Análisis Semántico Latente Probabilístico (PLSA) y fue descrito por Hofmann (1999). En contraste con su predecesor LSA, PLSA tiene un una

base sólida de inferencia estadística. Además, mientras LSA se deriva del álgebra lineal y realiza una SVD sobre la matriz términos-documento, PLSA es un modelo generativo de clases latentes que realiza una descomposición de probabilidad de esa matriz como una mezcla de distribuciones multinomiales condicionalmente independientes que se modelizan utilizando el algoritmo de maximización de expectativas (EM). Sin entrar en un tratamiento matemático completo del algoritmo, EM es un método para encontrar las estimaciones de parámetros más probables para un modelo que depende de variables latentes no observadas (en nuestro caso, los tópicos).

Una de las suposiciones de los modelos de tópicos es que cada documento consiste en una mezcla de tópicos, y cada tópico consiste en una colección de palabras. PLSA agrega un giro probabilístico a este supuestos:

- dado un documento d, el tema z está presente en ese documento con probabilidad $P(z/d)$;

- dado un tema z, la palabra w se extrae de z con probabilidad $P(w/z)$.

En la Figura 2 se muestra de forma gráfica el modelo con la notación de “placas”, que a menudo se usa para representar modelos gráficos probabilísticos. Los cuadros son "placas" que representan réplicas, que son entidades repetidas. La placa exterior representa documentos (M), mientras que la placa interior representa las posiciones de palabras repetidas en un documento (N) dado; cada posición está asociada con una elección de un tópico (z) y de una palabra (w). Los nodos se etiquetan utilizando la variable que representan, los nodos sombreados en gris representan las variables

aleatorias observadas en el modelo, mientras que los no sombreados representan variables aleatorias latentes en el modelo.

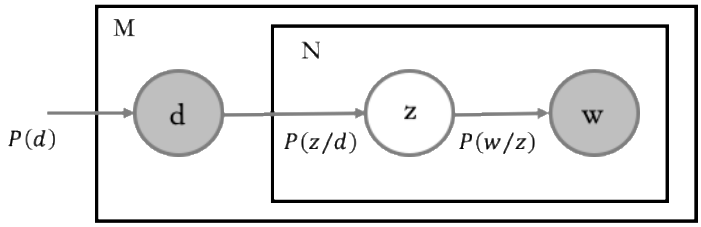

Figura 2. Notación de placa que representa el modelo asimétrico del Análisis Semántico Latente Probabilístico. (adoptado Blei *et al.*, 2003).

La probabilidad conjunta de observar una palabra y un documento dados es:

$$P(D,W) = P(D)\sum_Z P(Z/D)\,P(W/Z) \quad (3)$$

$$P(D,W) = \sum_Z P(Z)P(D/Z)\,P(W/Z) \quad (4)$$

La ecuación (3) indica cómo de probable es ver algún documento, y luego, según la distribución de tópicos de ese documento, cómo de probable es encontrar una cierta palabra dentro de ese documento. En la ecuación, $P(D)$, $P(Z/D)$ y $P(W/Z)$ son los parámetros del modelo. $P(D)$ puede ser determinado a partir del corpus mientras que los parámetros restantes son modelizados como distribuciones multinomiales utilizando el algoritmo EM.

La ecuación (4) es otra parametrización que utiliza un conjunto diferente de tres parámetros, en ella podemos ver un paralelismo directo entre los modelo PLSA y LSA donde:

$P(Z)$ =S, $P(D/Z)$= U y $P(W/Z)$= V^T. (5)

PLSA es un modelo mucho más flexible, pero aún tiene algunos problemas. En particular, como no tenemos parámetros para modelizar P(D), no sabemos cómo asignar probabilidades a nuevos documentos. Además, el número de parámetros para PLSA crece linealmente con el número de documentos que tenemos, por lo que es propenso a dar como resultado un modelo sobreajustado

2.4 ASIGNACIÓN LATENTE DE DIRICHLET (LDA)

La Asignación Latente de Dirichlet es una versión bayesiana de PLSA que se basa en el supuesto de la *bolsa de palabras* (Blei *et al.*, 2003). Esto significa que las palabras en un documento son intercambiables y los documentos se representan como secuencias de palabras individuales. Blei *et al.* (2003) aplicaron inicialmente el modelo LDA a corpus textuales, pero su uso se ha extendido también a imágenes (Iwata *et al.*, 2007) y videos (Wang *et al.*, 2007). En este documento nos restringiremos al análisis de una colección de textos científicos en el campo de la acuicultura.

El modelo LDA es un modelo generativo, es decir, un modelo que estudia cómo se producen los datos y una vez que se tiene el modelo de cómo se generan, se pregunta qué variable objetivo los ha generado. LDA postula que las características de los tópicos y documentos se extraen de la distribución Dirichlet, que es la generalización multivariante de la distribución beta.

Siguiendo a Blei *et al.* (2003), el modelo generativo puede resumirse como sigue:

1. Para cada tópico *k*, se extrae una distribución sobre las palabras $\phi_k \sim Dir(\alpha)$;

2. Para cada documento *d*,

   a) Se extrae un vector de proporción de tópicos $\theta_d \sim Dir(\beta)$;

   b) Para cada palabra *i*,

      i) Se extrae un tópico asignado $z_{d,i} \sim Mult(\theta_d), z_{d,n} \in \{1, \dots, K\}$;

      ii) Se extrae una palabra $w_{d,i} \sim Mult(\phi_{zd,i}), w_{d,i} \in \{1, \dots, V\}$.

En la Tabla 1 se describe la nomenclatura utilizada en la descripción del modelo generativo.

Tabla 1. Definición de las variables en el modelo LDA

Variable	Significado
$\boldsymbol{K}$	número de tópico
$\boldsymbol{V}$	número de palabras en el vocabulario
$\boldsymbol{M}$	número de documentos
$\boldsymbol{\mathcal{N}_{d=1\dots M}}$	número de palabras en el documento *d*
$\boldsymbol{N}$	el número total de palabras en todos los documentos; suma de todos los valores, es decir, $N_d N = \sum_{d=1}^{M} N_d$
$\boldsymbol{\alpha_{K=1\dots K}}$	hiperparámetro a priori del tópico *k* en un documento; por lo general el mismo para todos los temas; normalmente un número menor que 1, por ejemplo 0.1, a preferir distribuciones tema dispersos, es decir, pocos tópicos por documento
$\boldsymbol{\alpha}$	colección de todos los valores del hiperparámetro, visto como un único vector α_k
$\boldsymbol{\beta_{W=1\dots V}}$	hiperparámetro a priori de la palabra *w* en un tema; suele ser el mismo para todas las palabras; normalmente un número mucho menor que 1, por ejemplo, 0.001, a preferir fuertemente distribuciones de palabras dispersas, es decir, pocas palabras por tema
$\boldsymbol{\beta}$	colección de todos los valores del hiperparámetro, visto como un único vector β_w
$\boldsymbol{\varphi k = 1 \dots K, \omega = 1 \dots V}$	probabilidad de que palabra *w* se encuentren el en tópico *k*
$\boldsymbol{\varphi k = 1 \dots K}$	la distribución de las palabras en el tópico *k*
$\boldsymbol{\theta_{d=1\dots M,k=1\dots K}}$	probabilidad de tópico *k* se produzca en el documento *d*
$\boldsymbol{\theta_{d=1\dots M}}$	distribución de los tópicos en el documento *d*
$\boldsymbol{\mathcal{Z}_{d=1\dots M,k=1\dots \mathcal{N}_d}}$	identidad del tópico de la palabra *w* en el documento *d*
Z	identidad del tópico de todas las palabras en todos los documentos
$\boldsymbol{w_{d=1\dots M,w=1\dots \mathcal{N}_d}}$	la identidad de la palabra *w* en el documento *d*
W	la identidad de todas las palabras en todos los documentos

Fuente: Blei *et al.* (2003)

El objetivo con LDA es inferir o estimar las variables latentes, es decir, calcular su distribución condicionada a los documentos. Una representación gráfica en notación de placas del modelo LDA se encuentra en la Figura 3, en la que se puede deducir la siguiente distribución conjunta:

$$P(w,z,\theta,\phi/\alpha,\beta) = P(\theta/\alpha)\,P(z/\theta)\,P(\phi/\beta)\,P(w/z,\phi) \quad (6)$$

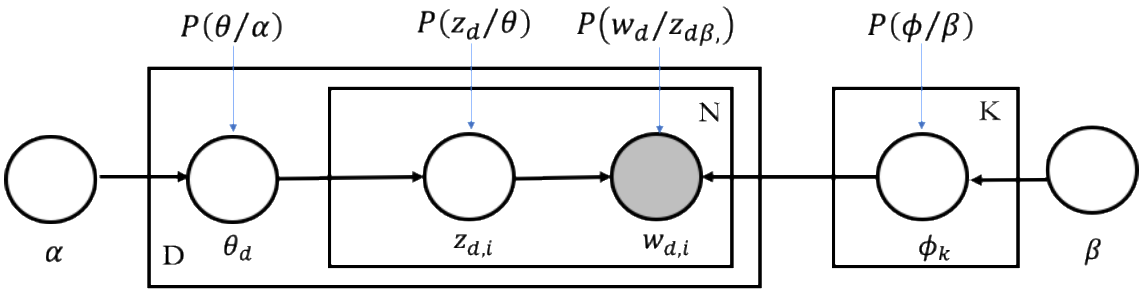

Figura 3. Notación de placa del modelo de Asignación Latente de Dirichlet. (adoptado Blei *et al.*, 2003).

En el lado derecho de la ecuación 6 encontramos:

- $P(\theta/\alpha)$: la distribución de tópicos por documento, dado el parámetro Dirichlet, que es un vector-K con componentes α_k > 0 (en la ecuación 7 se introduce el uso del operador de punto ($\alpha \cdot$) en el índice de variables como una abreviatura para la suma de todos los valores de las variables).

$$P(\theta/\alpha) = \frac{\Gamma(\alpha\cdot)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} \quad (7)$$

- $P(z/\theta)$: distribución del tópico z en el corpus, que depende de la distribución mencionada anteriormente. Por lo tanto, a cada palabra w_i en un documento de N palabras, se le asigna un valor de 1, …, K.

$$P(z/\theta) = \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{n_{d,k}} \quad (8)$$

En la distribución conjunta $P(z/\theta)$ se expresa la probabilidad de z para todos los documentos y tópicos en términos del número de palabras $n_{d,k}$, que es la cantidad de veces que se ha asignado el tema k a cualquier palabra en el documento d.

- $P(\phi/\beta)$: distribuciones de términos por tópico de todo el corpus ϕ_k. Se obtienen (nuevamente) de una distribución de Dirichlet con parámetro β. $\phi_{k,v}$ nos da la probabilidad de que el término v se obtenga cuando el tópico fuese elegido, se expresa para para todos los tópicos y todas las palabras del vocabulario como:

$$P(\phi/\beta) = \prod_{k=1}^{K} \frac{\Gamma(\beta_k, .)}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \prod_{v=1}^{V} \phi_{k,v}^{\beta_{k,v}-1} \quad (9)$$

- Finalmente, la probabilidad de un corpus w dado z y ϕ en el modelo gráfico es:

$$P(w/z, \phi) = \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{.,.,k,v}} \quad (10)$$

Se podría reescribir la ecuación 6 marginando las variables latentes con el fin de considerar un modelo de probabilidad dado, un corpus w y los hiperparámetros (α, β). Esta probabilidad es necesaria para poder realizar una "estimación máximo-verosímil de los parámetros del modelo y para inferir la distribución de las variables latentes (Blei *et al.,* 2003)

$$P(w/z,\phi) =$$

$$\int_{\phi}\int_{\theta}\sum_{z}\left(\prod_{d=1}^{D}\frac{\Gamma(\alpha\cdot)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k+n_{.,k,v}-1}\right)\left(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k,\cdot})}{\prod_{v=1}^{V}\Gamma(\beta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_{k,v}+n_{.,k,v}-1}\right)d\theta\, d\phi$$

(11)

Blei *et al.*, (2003) afirman que la suma sobre todas las combinaciones posibles de asignaciones de tópicos hace que esta probabilidad sea computacionalmente intratable y, por lo tanto, se tiene que hacer uso de algoritmos de *Machine Learning* para encontrar aproximaciones de la probabilidad marginal. Aunque la probabilidad a posteriori no puede calcularse exactamente, se puede lograr una aproximación lo suficientemente cercana al verdadero valor mediante inferencia estadística. En ese sentido podemos distinguir dos tipos de algoritmos inferenciales: (1) algoritmos basados en variaciones (por ejemplo, Blei y Jordan, 2006; Teh *et al.*, 2007; Wang *et al.,* 2011) y (2) los basados en muestreo (por ejemplo, Newman *et al.*, 2007; Porteous *et al.*, 2008). Los algoritmos basados en variación crean una familia de distribuciones más cercanas (la distancia se mide con la divergencia de Kullback-Leibler (KL)) al verdadero valor a posteriori. Es importante tener en cuenta que los algoritmos basados en variación y en muestreo proporcionan resultados igualmente precisos (Asunción *et al.*, 2012).

Es necesario medir el rendimiento de un modelo de tópicos para garantizar que se pueda generalizar a partir de los datos de una manera útil. Como se mencionó anteriormente, los modelos de tópicos son modelos de variables latentes que utilizan correlaciones entre las palabras y los tópicos semánticos latentes en una colección de documentos (Blei y Lafferty 2007). Esta definición supone entonces que el número esperado de tópicos (es decir, las variables latentes) debe establecerse a *priori*.

Consecuentemente, elegir el número adecuado de tópicos para una determinada colección de documentos no es trivial y siempre se debe buscar un balance entre la necesidad de una gran cantidad de tópicos (para cubrir todos los tópicos en la colección de documentos) y la necesidad de un número limitado de ellos que los expertos puedan comprender y verificar más fácilmente.

Dependiendo de los objetivos y de los medios computacionales disponibles, un investigador puede emplear una diferentes métricas de rendimiento (Wallach *et al.*, 2009; Buntine, 2009; Chang y Blei, 2009), Entre las medidas para evaluar el desempeño de los modelos de tópicos probabilísticos en minería de texto, lenguaje natural, procesamiento y muchas otras áreas de recuperación de información, se encuentran:

- La *perplejidad* definida (para un conjunto de prueba de M documentos) por Blei, *et al*. (2003) como:

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} \log P\,(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (10)$$

donde N_d es el número de palabras en el documento d-esímo del corpus de prueba D_{test} y w_d es el documento y w_d son las palabras el corpus. Un valor de perplejidad más bajo indica un mejor modelo.

- La validación cruzada. En algunos escenarios, el modelo puede evaluarse mediante validación cruzada del error de una tarea externa en cuestión, como la clasificación de documentos o la recuperación de información (Wei y Croft, 2006; Titov y McDonald, 2008).
- Verosimilitud empírica (Li y McCallum, 2006).
- Verosimilitud marginal, que puede ser aproximada por:
    - Media armónica (Newton and Raftery, 1994; Griths y Steyvers,2004)

- Chib-style (Chib ,1995)
- Left-to-right samplers, (Del Moral *et al*.,2006);

Además de las métricas anteriores, encontramos la *coherencia* (Röder *et al*., 2015), que se basa en la hipótesis de distribución (Harris, 1954), que afirma que las palabras con significados similares tienden a coexistir en contextos similares.

La coherencia utiliza cuatro etapas para llegar a una puntuación general para el tópico (Röder *et al*., 2015), a saber: (1) segmentación de las N palabras principales del tema en pares; (2) cálculo de la probabilidad de cada palabra individuales o pares de palabras; (3) cálculo de una medida de confirmación que captura el acuerdo de pares; y finalmente, (4) la agregación de medidas de confirmación individuales en una puntuación general de coherencia del tema.

2.5 COMPARACIÓN DE LOS MÓDELOS

En las secciones anteriores se realizó una descripción de los modelos LSA (Deerwester *et al*., (1990); Landauer *et al*., (1998)), PLSA (Hofmann, 1999) y LDA (Blei *et al*., 2003). En este apartado, se realizará una comparación resumida teniendo en cuenta las ventajas y desventajas de cada modelo (Tabla 2).

Como característica común de los tres modelos podemos anotar lo siguiente:

- Se basan en el mismo supuesto básico: cada documento consta de una mezcla de tópicos y cada tópico consiste en una colección de palabras;
- El número de tópicos (K) es un parámetro: ninguno de los algoritmos puede inferir K en la colección de documentos

Tabla 2. Ventajas y desventajas de los modelos LSA, PLSA y LDA.

	LSA	PLSA	LDA
Ventajas	-SVD en LSA solo ejecuta tratamiento matemático a la matriz, que no necesita gramática, semántica y otros conocimientos básicos del procesamiento de lenguaje natural (Girolami, M., & Kabán, A.,2003). - La dimensión espacial es reducido en gran medida haciendo que el problema sparce de datos mejore (Torkkola, K., 2002)	- Comparado con LSA, tiene una sólida base estadística (Brants, T.,2005) - Utiliza algoritmo de maximización de expectativas (EM) para entrenar clases latentes. y obtiene soluciones por iteración mientras computa el modelo, reduciendo en gran medida la complejidad del tiempo y aumentando la velocidad informática (Buntine, W., 2009)	- Modelo generativo no supervisado y y puede usar algoritmos de inferencia de probabilidad eficientes para calcular los parámetros del modelo (Blei *et al.*, 2003) - Es adecuado para manejar corpus de gran tamaño, puesto que el tamaño del espacio de parámetros del modelo LDA no tiene nada que ver con el número de documentos de entrenamiento (Lu, *et al.*, 2011) - El modelo LDA incorpora el hiperparámetro al nivel de tema de documento. Se agrega la información a priori, lo que significa que los parámetros pueden verse como variables aleatorias. Además, hace que LDA se convierta en un modelo jerárquico con estructura más estable, evitando el sobreajuste (Girolami, M., & Kabán, A.,2003).
Desventajas	- Lento para calcular la SVD con datos a gran escala y aplicaciones reales bajo la capacidad operativa Hofmann, T. 1999) - Es difícil determinar el valor de K en el algoritmo SVD. Este se determina mediante ecuaciones empíricas mediante la comparación de posibles opciones una por una (Hofmann, T. 1999) - Carece fundamentación estadística (motivación para PLSA) y asume que las palabras y los documentos forman un modelo gaussiano conjunto (Hofmann, T. 1999)	- EM de PLSA es un algoritmo completamente sin supervisión, por lo que su convergencia es lenta mientras el algoritmo itera (Lu, *et al.*, 2011) - No es un modelo generativo bien definido para documentos nuevos. El modelo PLSI necesita obtener una probabilidad previa, que solo se basa en conjunto de entrenamiento existente. Para el texto fuera del conjunto de entrenamiento, no hay probabilidad previa (Lu, *et al.*, 2011)	- No apto para corpus pequeños o con distribución normal (Chen, L., 2017) - Aunque LSA, PLSA y LDA pueden identificar efectivamente la relación de tema entre diferentes documentos. Sin embargo, dado que estos modelos semánticos no consideran ningún parámetro relacionado con el tiempo (Yuan *et al.*, 2013), no pueden resolver eficazmente problema relacionados con el tiempo, por ejemplo tendencias temporales de tópicos

- Los algoritmos tienen como entrada la matriz termino-documento;

- Generan dos matrices: la matriz tópico-término y la matriz tópico-documento.

2.6 LDA EN EL ENTORNO R

El primer software LDA fue publicado por David Blei en 2004 como lda-c e implementa inferencia variacional, que fue descrita por primera vez en Blei *et al.*, 2003. A la fecha existe una gran variedad de software de modelado de tópicos de código abierto y cerrado. Esta sección ofrece una visión general sobre las implementaciones en el software R que permiten además del modelado de tópicos, el manejo de datos textuales.

La red integral de archivos de R (Comprehensive R Archive Network-CRAN), por sus siglas en inglés, proporciona orientación sobre los paquetes que son relevantes para determinadas tareas. Allí podemos encontrar una lista de paquetes relacionados con el procesamiento natural del lenguaje. En total encontramos 59 paquetes relacionados con la minería de texto y modelado de tópicos, de los cuales ocho realizan el modelado de tópicos LDA (Tabla 3).

Tabla 3. Paquetes del software R para Asignación Latente Dirichlet

Paquete	Titulo	Cita
lda	Collapsed Gibbs Sampling Methods for Topic Models	Jonathan Chang (2015). lda: Collapsed Gibbs Sampling Methods for Topic Models. R package version 1.4.2. https://CRAN.R-project.org/package=lda
lda.svi	Fit Latent Dirichlet Allocation Models using Stochastic Variational Inference	Nicholas Erskine (2019). lda.svi: Fit Latent Dirichlet Allocation Models using Stochastic Variational Inference. R package version 0.1.0. https://CRAN.R-project.org/package=lda.svi
ldaPrototype	Prototype of Multiple Latent Dirichlet Allocation Runs	Jonas Rieger (2020). ldaPrototype: Prototype of Multiple Latent Dirichle Allocation Runs. R package version 0.1.1. https://CRAN.R-project.org/package=ldaPrototype
LDATS	Latent Dirichlet Allocation Coupled with Time Series Analyses	Juniper L. Simonis, Erica M. Christensen, David J. Harris, Renata M. Diaz, Hao Ye, Ethan P. White and S.K. Morgan Ernest (2020). LDATS: Latent Dirichlet Allocation Coupled with Time Series Analyses. R package version 0.2.7. https://CRAN.R-project.org/package=LDATS
ldatuning	Tuning of the Latent Dirichlet Allocation Models Parameters	Murzintcev Nikita (2019). ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. R package version 1.0.0. https://CRAN.R-project.org/package=ldatuning
LDAvis	Interactive Visualization of Topic Models	Carson Sievert and Kenny Shirley (2015). LDAvis: Interactive Visualization of Topic Models. R package version 0.3.2. https://CRAN.R-project.org/package=LDAvis
topicdoc	Topic-Specific Diagnostics for LDA and CTM Topic Models	Doug Friedman (2019). topicdoc: Topic-Specific Diagnostics for LDA and CTM Topic Models. R package version 0.1.0. https://CRAN.R-project.org/package=topicdoc
topicmodels	Topic Models	Grün B, Hornik K (2011). "topicmodels: An R Package for Fitting Topic Models." _Journal of Statistical Software_, *40*(13), 1-30. doi: 10.18637/jss.v040.i13 (URL: https://doi.org/10.18637/jss.v040.i13).

3 OBJETIVOS

3.1 OBJETIVO GENERAL

Poner de manifiesto que la combinación de métodos de machine learning no supervisados y de métodos multivariantes clásicos permiten enriquecer la revisión exploratoria de literatura científica con un enfoque práctico del modelado probabilistico de tópicos.

3.2 OBJETIVOS ESPECIFICOS

- Analizar la literatura publicada en el campo de la acuicultura (para las revistas con más índice de impacto) durante los últimos 47 años, desde 1972 hasta 2019, desde una perspectiva probabilística bayesiana y conocer su evolución mediante la utilización de Biplots dinámicos.
- Identificar los principales tópicos en el área de la acuicultura en el periodo de tiempo evaluado;
- Describir patrones y tendencias temporales de los tópicos en acuicultura
- Describir patrones y tendencias de los temas dentro de las revistas top de acuicultura en el tiempo evaluado mediante análisis de datos de tres vías.

4 MÉTODOS

El procedimiento para la identificación de tópicos a través de LDA se dividió en cuatro etapas: (1) Búsqueda y recopilación de artículos, (2) Preprocesamiento, (3) Construcción del modelo LDA y (4) Etiquetado de tópicos.

4.1 BÚSQUEDA Y RECOPILACIÓN DE ARTÍCULOS

Como fuente de información para nuestra investigación, utilizamos la base de datos *Web of Science* (WOS) producida por *Thomson Reuters*. Entre las razones que nos motivaron a seleccionar esta base de datos, es que esta es una de las bases de datos más utilizadas por los investigadores (Harzing y Alakangas, 2016) y también, porque esta es una plataforma cuyo alto grado de impacto y reconocimiento es unánime en todas las áreas de investigación (Bar-Ilan, 2008).

La acuicultura no se considera como una categoría separada en la base WOS, sino que se encuentra dentro de la categoría "Fishing" que es definida en los siguientes términos: "La pesca abarca aquellos recursos relacionados con numerosos aspectos de la ciencia, la tecnología e industria pesquera, incluida la patología, la fisiología y la bioquímica de los peces, enfermedades de los peces y acuicultura". Teniendo en cuenta lo anterior, incluimos en la busqueda todas las revistas con el término "Aquaculture" o "Aquacultural" en la categoría "Fishing" con un factor de impacto mayor o igual a 1.0. También incluimos algunas revistas que, aunque no incluyen estas palabras, abordan

explícitamente tópicos de acuicultura. Sólo se tuvieron en cuenta resúmenes de artículos en idioma inglés y cada uno de ellos se descargó en formato "*.csv*" con tres metavariables: título, año de publicación y nombre de la revista. De esta forma conseguimos el corpus, es decir, la colección de resúmenes con los que trabajamos en las fases posteriores. Cabe señalar que este estudio tiene la limitación de que tanto libros, como reseñas, literatura gris e informes han sido excluidos de la investigación.

4.2 PREPROCESAMIENTO

El preprocesamiento de la información juega un papel muy importante y generalmente es el primer paso en técnicas y aplicaciones de minería de textos (Vijayaran *et al*., 2015). Los métodos de preprocesamiento tienen como objetivo eliminar el ruido o los datos sin sentido de un corpus. En esta fase, los resúmenes de los artículos se transformaron a un formato legible para el algoritmo que se aplicó en una fase posterior. Los pasos siguientes de preprocesamiento de minería de texto estándar se aplicaron en todo el corpus:

(i) Para aumentar la coherencia de los tópicos, cada resumen se dividió en n-grama que en el campos de la lingüística computacional es una secuencia contigua de n elementos (en nuestro caso palabras) de una muestra dada de texto o discurso (Jurasky, D., & Martin, J. H., 2000);. Con el fin de que fueran preservados los nombres científicos de las especies fueran preservadas, por ejemplo, "*Oreochromis niloticus*", "*Dicentrarchus labrax*", "*Sparus aurata*", entre otros se utilizaron bigrama.

(ii) Eliminación de palabras denominadas *stopword* (usualmente artículos, pronombres, preposiciones y conjunciones) que son palabras que no tienen significado léxico. Aparecen en los textos con mucha frecuencia, no contiene información que sea de interés para el análisis, perjudica la precisión de los resultados y además reduce la dimensionalidad del espacio de tópicos. En este caso utilizamos el método clásico de eliminación el cual, según Vijayaran *et al.* (2015), se basa en la eliminación de una lista precompilada;

(iii) Poner en minúscula todo el texto, para evitar que una palabra se cuente dos veces debido a la capitalización;

(iv) Eliminación de números, caracteres de puntuación y espacios en blanco.

Se consideró que la interpretación de los tópicos es un aspecto muy importante a ser tenido en cuenta en una fase posterior, por lo que se omitió la fase de normalización comúnmente utilizada (*stemming* y *lemmantization*) dado que puede reducir la capacidad de interpretación, puesto que los algoritmos utilizados pueden ser demasiado agresivos y pueden dar lugar a palabras irreconocibles. Además, puede conducir a un problema de identificación ya que una forma derivada de una palabra podría provenir de un verbo o de un sustantivo (Evangelopoulos *et al.,* 2012).

Como resultado de la fase de preprocesamiento se creó la matriz de términos de documentos (dtm).

Los pasos del preprocesamiento se realizaron con la ayuda del paquete de datos textuales *textmineR* (Jones, 2019) del software R (Core Team, 2019).

4.3 CONSTRUCCIÓN DEL MODELO LDA

Como se mencionó anteriormente el número óptimo de tópicos no se conoce a *priori*, por lo que se probaron 12 diferentes modelos variando el número de tópicos (K) de 5 a 60 con incrementos de 5, realizando 1000 iteraciones para el muestreo de Gibbs (Geman y Geman, 1984); también, para los valores de parámetros de Dirichlet, α se ajustó mediante optimización del paquete *textmineR* mientras que ß fue estimada a partir del corpus. Como métrica para la selección del mejor modelo se utilizó la coherencia. La Tabla 4 muestra la configuración experimental utilizada.

Tabla 4. Configuración experimental de parámetros utilizados en la creación del modelo Asignación Latente de Dirichlet.

Componente	Candidato
Algoritmo de inferencia	Gibbs sampling
El número de tópicos	5,10,15,20,25,30,35,40,45,50,55,60
Número de iteraciones	1000
Parámetro Dirichlet α	Optimizado por el paquete textmineR
Parámetro Dirichlet β	Estimado a partir del corpus

4.4 ETIQUETADO DE TÓPICOS

Los tópicos no están semánticamente etiquetados para el modelo LDA. Lewis *et al.* (2013) mencionan que los análisis algorítmicos tienen una capacidad muy limitada para comprender los significados latentes del lenguaje humano, por lo que el etiquetado manual se considera un estándar (Lau *et al.*, 2011). Para proporcionar una interpretación semánticamente correcta, un experto en el dominio de la acuicultura, con más de diez años de experiencia, etiquetó manualmente los tópicos utilizando las siguientes fuentes de información: una lista de las 10 palabras más frecuentes en cada tópico y una muestra de tres títulos y resúmenes de artículos pertenecientes a cada tópico. Además, para confirmar el etiquetado de los tópicos, los visualizamos en un área bidimensional calculando la distancia entre los tópicos (Chuang *et al.*, 2012), en dicha visualización los centros de los círculos que representan a los tópicos se presentan de acuerdo con un algoritmo de escalamiento multidimensional (MDS). Utilizamos la divergencia de Jensen-Shannon para calcular las distancias entre los tópicos utilizando la funcion LDAvis (Siever & Shirley, 2014).

4.5 INDICES CUANTITATIVOS

Puesto que no es posible determinar los patrones y tendencias temporales de los tópicos de forma intiuitiva, se utilizaron algunos índices propuestos por Xiong *et al.* (2019), los cuales se obtienen a partir de las distribuciones de probabilidad de palabras por tópico y la distribución de tópicos por documento y se describen como sigue:

La distribución de tópicos a través del tiempo se obtiene como:

$$\theta_k^y = \frac{\sum_{m \epsilon y} \theta_{mk}}{n^y} \quad (11)$$

donde $m \epsilon y$ representa los artículos publicados en cierto año y, θ_{mk} la proporción del k-esímo tópico y n^y el número total de artículos publicados en el año y.

Con el fin de facilitar la caracterización de los tópicos en términos de su tendencia, se utilizó la pendiente de un modelo de regresión lineal simple, donde los años fueron a variable independiente y la proporción de los tópicos fue la variable respuesta (Griffiths & Steyvers, 2004). Los tópicos cuyas pendientes de regresión fueron positivas (negativas) a un nivel de significación estadística de 0.01 se clasificaron como tópicos crecientes (decrecientes) y en caso contrario, es decir, cuando no hubo significación se clasificaron como fluctuantes.

La distribución de tópicos dentro de las revistas θ_k^j se definió como :

$$\theta_k^j = \frac{\sum_{m \epsilon j} \theta_{mk}}{n^j} \quad (12)$$

donde $m \epsilon j$ representan los artículos publicados en una revista particular *j*, θ_{mk} la proporción del k-esímo tópico y n^j el número total de artículos publicados en la revista *j*.

También, para cada revista se definió $\theta_k^{j,y}$ como la proporción de tópicos por año dentro de cada revista:

$$\theta_k^{j,y} = \frac{\sum_{m \epsilon j \cap m \epsilon y} \theta_{mk}}{n^{j,y}} \quad (13)$$

Para explorar si en una revista específica la cobertura de los tópicos es estrecha o amplia, es decir…. (completar), utilizamos la entropía e^j.

$$e^j = -\sum_{k=1}^{K} \theta_k^j \ln(\theta_k^j) \quad (14)$$

Un e^jgrande indica una distribución de tópicos más uniforme en una revista, mientras que en caso contrario indica que hay sido pocos los tópicos que fueron el foco de los estudios publicados en esa revista.

Se realizó un análisis de clúster utilizando el método jerárquico de medias aritméticas no ponderadas (UPGMA), (Sneath y Sokal, 1973) utilizando la distancia euclídea entre revistas. Una distancia menor entre dos grupos sugiere un mayor grado de similitud entre ellos, es decir, las revistas de cada grupo tienen contenidos e intereses similares.

Para cuantificar la actividad y la influencia de los tópicos, calculamos su popularidad cuya métrica tiene en cuenta tanto la tendencia como la probabilidad de aparición.

$$P^i = S_{NP}^i + S_{Tr}^i \quad (15)$$

$$S_{NP}^i = P_A^i / P_A^{max} \quad (16)$$

donde P^i es la popularidad del tópico *i*, S_{NP}^i es la probabilidad normalizada y S_{Tr}^i toma valores de 1, 0.67 and 0.33 si un tópico muestra tendencia positiva, fluctuante o negativa, respectivamente.

4.6 ANÁLISIS BIPLOT

Siguiendo un enfoque similar al análisis de las tendencias de los tópicos, se agruparon los tópicos en temas generales. Los datos así agrupadaos se analizaron mediante un Biplot Dinámico (Egido y Galindo, 2015), que es una técnica que se utiliza cuando se

quiere analizar un conjunto de datos de tres vías (en nuestro caso revistas en filas, temas en columnas y años para las diversas ocasiones o situaciones en la tercera vía). Se utilizaron los siguientes intervalos de tiempo, 2000–2004, 2005–2009, 2010–2014 y 2015-2019. Se fijó como referencia el último quinquenio.

Siguiendo a Egido y Galindo (2015), el cálculo del Biplot Dinámico se realiza en dos etapas:

Etapa 1. Análisis estático. En esta etapa se realiza un análisis Biplot del período referencia, es decir, los datos correspondientes al último quinquenio. Si bien es cierto que en esta etapa se podría utilizar cualquiera de las factorizaciones Biplot propuestas por Gabriel en 1971, JK-Biplot o GH-Biplot, se siguió la recomendación de Ejido y Galindo (2015) quienes consideran que la más apropiada es la correspondiente al análisis HJ-Biplot (Galindo, 1986). Este biplot es una técnica definida para la representación conjunta, en un espacio vectorial de baja dimensión (generalmente dos), de las filas y columnas de una matriz de datos, utilizando marcadores (puntos / vectores), para sus filas y para sus columnas (respectivamente). Al igual que los biplots clásicos propuestos por Gabriel (1971), la interpretación se basa en conceptos geométricos simples: (i) las distancias entre los marcadores de fila (revistas): se interpretan como una función inversa de su similitud, de tal manera que los marcadores más cercanos representan a revistas que son más similares; (ii) las longitudes de los marcadores (vectores) de columna (temas) se aproximan a la desviación estándar de las variables (a la variabilidad de los temas); (iii) los cosenos de los ángulos entre los vectores columna aproximan las correlaciones entre las variables, de tal manera que los ángulos agudos pequeños se asocian con variables con correlaciones positivas altas,

los ángulos obtusos con correlaciones negativas también altas, mientras que ángulos rectos se asocian con variables no correlacionadas. De la misma manera, los cosenos de los ángulos entre los marcadores variables y los ejes (Componentes Principales) se aproximan a las correlaciones entre ellos. (iv) El orden de las proyecciones ortogonales de los marcadores fila sobre el vector que representa a una columna (tema), se aproxima al orden de los elementos de fila (revistas) en esa columna, de manera que aquellas revistas que se proyectan cercanas o más allá a la punta de la flecha que representa a un tema, significa que ese tema ha sido importante en esa revista, y viceversa.

Etapa 2. El objetivo en esta etapa consiste en proyectar, sobre el biplot obtenido en el paso previo (biplot referencia), el resto de situaciones, en decir, el resto de los quinquenios considerados en el estudio. De esta forma, se pueden interpretar la dinámica de la evolución o trayectoria, de cada revista, enriqueciendo notablemente la interpretación gráfica.

Los elementos proyectados en diferentes tiempos o situaciones conservan propiedades similares a la de la factorización elegida con respecto a la situación de referencia (quinquenio 2014-2019).

La parte operativa de cálculos y gráficos se desarrolló con el paquete del software R dynBiplotGUI versión 1.1.5 (Ejido, 2017).

5 RESULTADOS Y DISCUSIÓN

Teniendo en cuenta los criterios de inclusión, se consideraron aptos para el análisis 38319 resúmenes de artículos pertenecientes a 14 revistas publicadas en el período entre 1972 y 2019. La Figura 4 muestra la distribución del número de artículos por revista y año, mientras que en la Tabla 5 se encuentra el nombre, abreviatura, el ranking y factor de impacto de cada revista, el rango de años en el cual se recolectaron los artículos, y además, el número medio de palabras encontradas y el vocabulario (palabras únicas). El período inicialmente seleccionado de 47 años, no se cubrió totalmente por todas las revistas, (por ejemplo, *Reviews in Aquaculture* comenzó en 2009), por lo tanto, para esta revista, solo se incluyeron en el estudio los artículos de ese año en adelante.

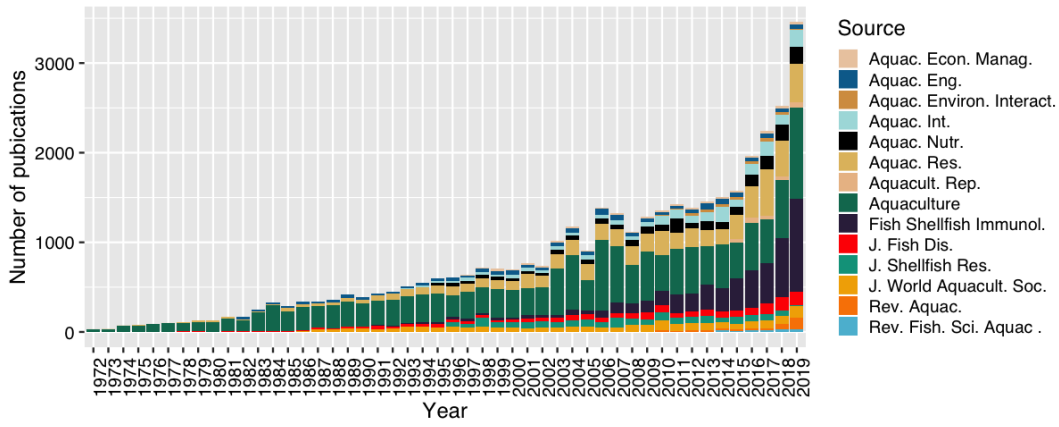

Figura 4. Número de publicaciones por revista entre los años 1972-2019 que se utilizaron para crear el modelo de Asignación Latente de Dirichlet. El número total de documentos fue de 38319.

Tabla 5. Descripción general del conjunto de datos utilizado para identificar y analizar la tendencia de tópicos en acuicultura con el modelo de Asignación Latente de Dirichlet. El ranking y el factor de impacto se extrajeron de los Informes de citas ISI 2018 (JCR) proporcionados por Thomson Reuters. N es el número de resúmenes que se consideran

aptos para su posterior análisis. $\overline{W}$ es el promedio de palabras, Ds W desviación estándar estimada del número de palabras, y $\overline{V}$ es el tamaño medio del vocabulario.

Nombre de la Revista	Abreviatura	Ranking Fisheries-Science	Factor de impacto	N	Rango de años	$\overline{W}$	Ds W	$\overline{V}$
Reviews in Aquaculture	Rev. Aquac.	1	7,190	327	2009-2019	255	57	143
Reviews in Fisheries Science and Aquaculture	Rev. Fish. Sci. Aquac.	3	3,775	139	2014-2019	255	70	144
Fish and Shellfish Immunology	Fish Shellfish Immunol.	6	3,298	4893	1991-2019	301	83	150
Aquaculture Economics and Management	Aquac. Econ. Manag.	7	3,250	448	1997-2019	191	53	111
Aquaculture	Aquaculture	9	3,022	15805	1972-2019	292	114	141
Aquaculture Environment Interactions	Aquac. Environ. Interact.	15	2,380	244	2010-2019	286	46	150
Aquacultural Engineering	Aquac. Eng.	17	2,143	1346	1982-2019	263	105	136
Aquaculture Nutrition	Aquac. Nutr.	18	2,098	1845	1995-2019	291	70	137
Journal of Fish Diseases	J. Fish Dis.	20	1,988	1811	1978-2019	235	61	127
Aquaculture Reports	Aquacult. Rep.	21	1,887	230	2015-2019	302	71	147
Aquaculture Research	Aquac. Res.	30	1,502	5745	1972-2019	268	72	133
Aquaculture International	Aquac. Int.	32	1,455	1826	1993-2019	283	79	139
Journal of the World Aquaculture Society	J. World Aquacult. Soc.	33	1,386	2043	1986-2019	277	87	138
Journal of Shellfish Research	J. Shellfish Res.	35	1,037	1617	1995-2019	290	97	144

Los términos de mayor frecuencia encontrados en todos los documentos analizados se muestran en la Figura 5, representados mediante una nube de palabras (o *wordcloud*) y un diagrama de barras. La primera es una representación que visualmente puede ser más atractiva que el diagrama de barras, especialmente si el número de términos es grande. En la nube de palabras los tamaños de las palabras son proporcionales a las probabilidades de ocurrencia, de manera que esta forma de visualizar gráficamente la información estadística nos permite una visión general de la variedad de temas investigados en el campo de la acuicultura.

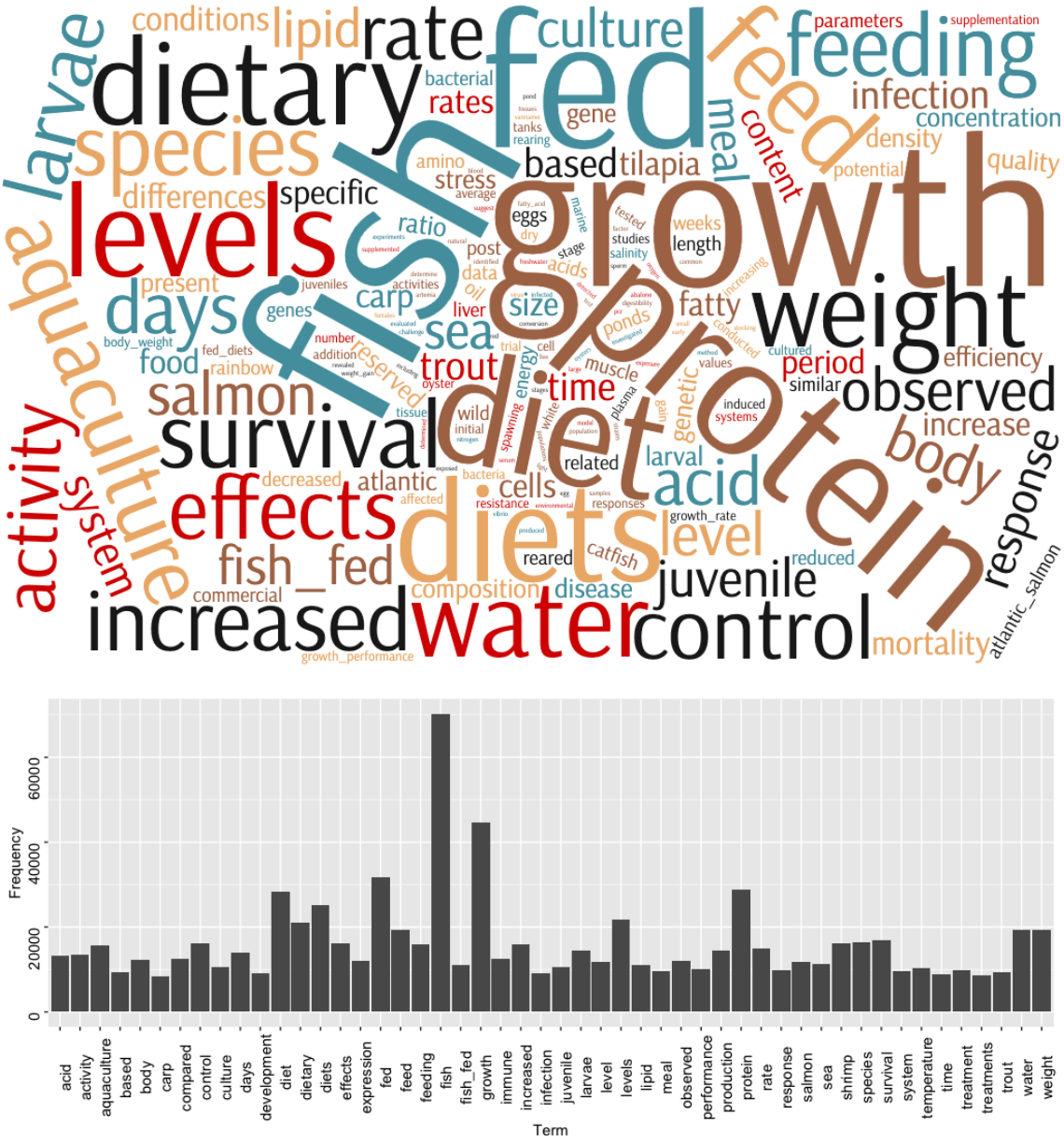

Figura 5. Nube de palabra (180 palabras) y gráfico de barras (50 palabras) de los términos más frecuentes en el corpus completo.

En la Tabla 6 se muestra los bigramas más frecuentes referidos a las especies encontradas en el corpus, el número de documentos (Fd) en los cuales aparece mencionada dicha especie y la frecuencia de documento inversa (IDF).

Tabla 6. Bigramas más frecuentes de especies en el corpus total.

Bi-grama	F	Fd	idf	Bi-grama	F	Fd	IDF
rainbow_trout	6155	2436	2,76	chinook_salmon	483	205	5,23
atlantic_salmon	5511	2395	2,77	seabream_sparus	467	319	4,79
salmo_salar	2735	2009	2,95	yellow_croaker	454	167	5,44

nile_tilapia	2716	1099	3,55	crassostrea_virginica	442	285	4,90
salmon_salmo	2621	1930	2,99	penaeus_vannamei	439	268	4,96
litopenaeus_vannamei	2538	1430	3,29	salmo_trutta	401	268	4,96
oncorhynchus_mykiss	2321	1648	3,15	epinephelus_coioides	401	243	5,06
trout_oncorhynchus	2265	1618	3,16	haliotis_discus	387	221	5,16
oreochromis_niloticus	2241	1424	3,29	eriocheir_sinensis	387	208	5,22
tilapia_oreochromis	2154	1456	3,27	mytilus_edulis	386	259	5,00
sea_bass	1972	698	4,01	mitten_crab	385	180	5,36
grass_carp	1971	569	4,21	oysters_crassostrea	382	312	4,81
common_carp	1854	797	3,87	sturgeon_acipenser	378	245	5,05
sea_bream	1840	682	4,03	bluefin_tuna	378	152	5,53
white_shrimp	1521	873	3,78	oyster_pinctada	376	233	5,10
shrimp_litopenaeus	1492	1010	3,64	salmo_gairdneri	368	297	4,86
cyprinus_carpio	1396	948	3,70	crucian_carp	368	126	5,72
penaeus_monodon	1365	790	3,88	gibel_carp	360	104	5,91
carp_cyprinus	1206	851	3,81	hippoglossus_hippoglossus	343	231	5,11
macrobrachium_rosenbergii	1078	596	4,16	mussel_mytilus	340	260	4,99
sparus_aurata	1068	702	4,00	halibut_hippoglossus	335	224	5,14
abalone_haliotis	1043	642	4,09	brachionus_plicatilis	326	234	5,10
sea_cucumber	1041	371	4,64	penaeid_shrimp	322	233	5,10
crassostrea_gigas	1018	626	4,11	carp_carassius	291	190	5,31
shrimp_penaeus	963	665	4,05	red_tilapia	289	112	5,84
artemia_nauplii	948	519	4,30	pacific_oysters	288	178	5,37
dicentrarchus_labrax	884	607	4,15	procambarus_clarkii	285	171	5,41
oyster_crassostrea	868	622	4,12	megalobrama_amblycephala	283	160	5,48
ictalurus_punctatus	853	598	4,16	salvelinus_alpinus	282	175	5,39
atlantic_cod	804	386	4,60	cherax_quadricarinatus	277	148	5,56
grouper_epinephelus	775	478	4,38	vannamei_boone	268	216	5,18
prawn_macrobrachium	720	487	4,37	isochrysis_galbana	260	234	5,10
trout_salmo	709	523	4,29	crab_scylla	250	152	5,53
carp_ctenopharyngodon	672	437	4,47	ostrea_edulis	249	152	5,53
bass_dicentrarchus	656	477	4,39	oncorhynchus_kisutch	246	178	5,37
sea_cucumbers	620	210	5,21	silver_perch	235	59	6,48
pearl_oyster	579	280	4,92	pagrus_major	234	155	5,51
ctenopharyngodon_idella	572	361	4,66	mussels_mytilus	216	176	5,38
clarias_gariepinus	546	308	4,82	oncorhynchus_tshawytscha	216	154	5,52
sea_urchin	529	251	5,03	scallop_argopecten	216	141	5,60
coho_salmon	528	228	5,12	portunus_trituberculatus	216	118	5,78
bream_sparus	521	371	4,64	tuna_thunnus	215	151	5,54
silver_carp	514	165	5,45	salmon_smolts	205	155	5,51
atlantic_halibut	513	235	5,09	rohu_labeo	205	148	5,56
sea_urchins	512	182	5,35	asian_seabass	204	79	6,18

F= Frecuencia del termino; Fd = Número de documentos ; idf= Frecuencia de documento invers

Se observa que aparecen nombres comunes y científicos que pertenecen a la misma especie, tales como trucha arcoiris (rainbow_trout) cuyo nombre científico es *Oncorhynchus mykiss* ; salmon del atlantico (atlantic_salmon), *Salmo salar* ; tilapia del nilo (nile_tilapia), *Oreochromis niloticus* ; camarón blanco (white_shrimp) *Litopenaeus vannamei* ; carpa común (common_carp) , *Cyprinus carpio*

5.1 IDENTIFICACIÓN DE LOS TÓPICOS

Dada la configuración experimental utilizada, encontramos que los resultados sugieren que el modelo LDA, con la puntuación de coherencia óptima, contiene 40 tópicos (k = 40) (Figura 6).

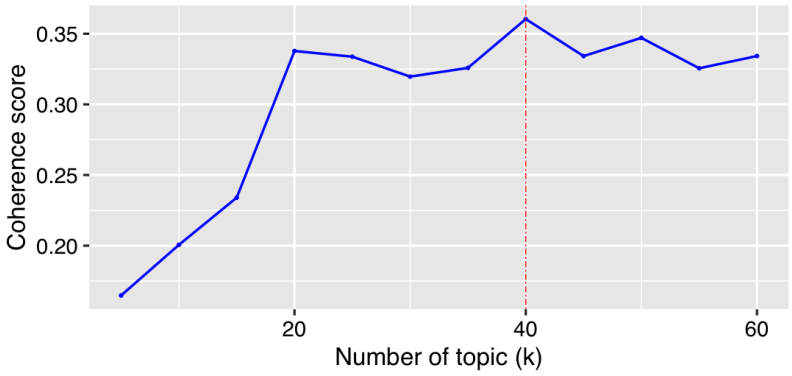

Figura 6. Puntuaciones de coherencia calculadas para el número de tópicos (k). El modelo de 40 tópicos presenta la mayor coherencia.

La Tabla 7 presenta los 40 tópicos estimados por nuestro modelo y, para cada uno de ellos, los 20 términos más frecuentes. Esta tabla también incluye la clasificación de los tópicos dentro de una lista de asignación fraccional en orden descendente de proporciones a lo largo de la colección de artículos. Del mismo modo, la tabla también presenta el número de artículos en los que cada tópico tiene la mayor proporción, es

decir, una asignación discreta. Cabe señalar que la diferencia entre entre asignaciones fraccionales y discretas se observa notablemente. Esto significa que existen diferencias en la clasificación de tópicos entre los dos tipos de clasificación. Por ejemplo, el tópico 2 (“diet”) ocupó el segundo lugar en términos de proporción, mientras que su clasificación en términos del número de artículos asignados se ubica en el décimo lugar. Esto indica que "diet" se ha investigado empíricamente en varios artículos, pero el tópico no se elige en muchos casos porque esos artículos también abarcan otros temas con mayores proporciones.

Los tópicos fueron manualmente etiquetados con una descripción lógica que mejor captura la semántica de las veinte palabras principales con la mayor probabilidad.

Tabla 7. Tópicos encontrados de los 38819 resúmenes de artículos de acuicultura publicados en el período 1972-2019 en 14 revistas especializadas en acuicultura.

Tópicos	Top-20 terminos	Share (Ranking)		Temas
		Proporcion(%)	Número de resumenes	
t_1Aquaculture production	aquaculture, production, species, fish, development, research, management, farming, potential, industry, economic, environmental, culture, marine, systems, based, review, commercial, information, studies	6.3 (1)	2718 (1)	Aquaculture production
t_2 Diet	fed, diet, diets, fish, dietary, fish_fed, control, growth, supplemented, supplementation, fed_diets, feeding, fed_diet, performance, effects, vitamin, levels, weeks, compared, growth_performance	4.5 (2)	1497 (10)	Nutrition
t_3 Growth performance	growth, fish, feed, weight, body, fed, performance, ratio, gain, efficiency, dietary, weight_gain, rate, conversion, growth_performance, fish_fed, juvenile, feed_conversion, growth_rate, body_weight	4.4 (3)	1028 (15)	Fish performance
t_4 Recirculating aquaculture system	water, system, systems, nitrogen, quality, culture, water_quality, aquaculture, production, oxygen, organic, concentrations, ammonia, carbon, dissolved, concentration, recirculating, treatment, flow, removal	4.0 (4)	2004 (3)	Husbandry protocols
t_5 Reproduction	eggs, spawning, females, egg, female, males, sex, reproductive, maturation, male, broodstock,	3.9 (5)	2295 (2)	Reproduction

Tópicos	Top-20 terminos	Share (Ranking)		Temas
		Proporcion(%)	Número de resumenes	
	development, triploid, hatching, stage, fertilization, induced, ovarian, hormone, production			
t_6 Molecular studies	expression, genes, immune, gene, response, related, infection, grouper, gene_expression, regulated, regulation, proteins, fish, cells, pathway, induced, immune_response, protein, responses, involved	3.7 (6)	1536 (9)	Genetic
t_7 Immune genetic response	expression, protein, amino, sequence, cdna, gene, tissues, domain, expressed, molecular, immune, role, acid, mrna, binding, peptide, length, amino_acid, characterization, acids	3.7 (7)	1730 (5)	Genetic
t_8 Larviculture and live feeds	larvae, larval, artemia, survival, feeding, live, development, nauplii, stage, rotifers, days, fed, stages, rearing, post, early, food, growth, dph, hatching	3.6 (8)	1950 (4)	Husbandry protocols
t_9 Diet formulation	meal, protein, diets, fish, diet, digestibility, fish_meal, soybean, feed, plant, replacement, soybean_meal, apparent, fishmeal, ingredients, fed, inclusion, based, sources, energy	3.5 (9)	1542 (8)	Nutrition
t_10 Disease	infection, disease, fish, virus, infected, mortality, pcr, viral, parasite, necrosis, detection, infections, host, infectious, detected, prevalence, samples, gill, caused, parasites	3.2 (10)	1566 (7)	Health
t_11 Bivalves	oyster, oysters, shell, crassostrea, gigas, scallop, clam, spat, clams, pearl, scallops, mortality, pacific, species, crassostrea_gigas, sites, summer, oyster_crassostrea, culture, bivalve	3.1 (11)	1582 (6)	Specific aquatic organisms
t_12 Immunity and disease	immune, activity, carp, resistance, lysozyme, control, parameters, serum, response, hydrophila, immune_response, increased, days, disease, immunity, responses, disease_resistance, challenge, innate, specific	3.0 (12)	1167 (11)	Health
t_13 Diet composition	protein, dietary, lipid, levels, diets, energy, level, fed, content, diet, dietary_protein, carbohydrate, composition, body, increased, growth, ratio, crude, protein_lipid, starch	2.9 (13)	954 (19)	Nutrition
t_14 Growth survival	growth, survival, rate, rates, growth_rate, weight, days, growth_survival, growth_rates, experiment, treatments, juvenile, juveniles, performance, differences, culture, length, control, effects, treatment	2.9 (14)	536 (29)	Fish performance
t_15 Modeling	model, method, fish, data, methods, based, time, system, models, developed, values, size, design, technique, studies, conditions, test, linear, flow, reserved	2.8 (15)	956 (18)	Fish performance

Tópicos	Top-20 terminos	Share (Ranking)		Temas
		Proporcion(%)	Número de resumenes	
t_16 Physiological responses	stress, fish, levels, plasma, exposure, exposed, turbot, blood, cortisol, glucose, response, effects, concentrations, increased, ammonia, physiological, sturgeon, control, concentration, responses	2.7 (16)	952 (20)	Health
t_17 Feed intake	feeding, feed, food, fed, growth, energy, consumption, rate, weight, intake, rates, fish, ration, daily, size, days, sea, body, urchins, period	2.5 (17)	1000 (16)	Nutrition
t_18 Intestinal_microbiota	bacterial, bacteria, vibrio, strains, strain, probiotic, isolates, isolated, fish, microbiota, resistance, probiotics, pathogenic, intestinal, aeromonas, microbial, pathogens, harveyi, parahaemolyticus, bacillus	2.4 (18)	1151 (13)	Health
t_19 Salmon	salmon, atlantic, atlantic_salmon, fish, salmo, salar, salmo_salar, salmon_salmo, cod, sea, lice, smolts, farmed, smolt, wild, atlantic_cod, parr, sea_lice, salmonis, seawater	2.3 (19)	970 (17)	Specific aquatic organisms
t_20 Genetic variability	genetic, populations, wild, species, population, markers, dna, loci, hybrids, microsatellite, hybrid, stocks, diversity, variation, individuals, strains, hatchery, gene, number, marker	2.3 (20)	1157 (12)	Genetic
t_21 Stocking density	density, stocking, size, densities, stocking_density, reared, fish, tanks, survival, juveniles, rearing, culture, hatchery, juvenile, small, tank, stocking_densities, production, stocked, large	2.3 (21)	807 (23)	Husbandry protocols
t_22 Sperm cryopreservation	sperm, treatment, min, concentrations, concentration, treated, motility, time, dose, control, treatments, storage, days, fertilization, spermatozoa, post, exposure, effective, semen, cryopreservation	2.3 (22)	1089 (14)	Husbandry protocols
t_23 Temperature and salinity tolerance	temperature, water, salinity, temperatures, water_temperature, seawater, ppt, conditions, salinities, na, effects, tolerance, increased, reared, cold, thermal, survival, range, temperature_salinity, atpase	2.2 (23)	680 (26)	Husbandry protocols
t_24 Fatty acids	fatty, acid, fatty_acid, acids, lipid, fatty_acids, composition, dha, content, acid_composition, dietary, pufa, lipids, levels, hufa, diets, polyunsaturated, profile, polyunsaturated_fatty, fed	2.1 (24)	904 (21)	Nutrition
t_25 Shrimp	shrimp, vannamei, catfish, white, wssv, litopenaeus, penaeus, litopenaeus_vannamei, channel, monodon, channel_catfish, white_shrimp, shrimp_litopenaeus, penaeus_monodon, shrimp_fed, shrimps, spot, white_spot, syndrome, shrimp_penaeus	2.1 (25)	389 (32)	Specific aquatic organisms
t_26 Immunization vaccines	fish, flounder, vaccine, antibody, vaccination, injection, protection, vaccinated, paralichthys, serum, tarda, specific, challenge, antibodies, japanese, olivaceus,	2.0 (26)	878 (22)	Health

Tópicos	Top-20 terminos	Share (Ranking)		Temas
		Proporcion(%)	Número de resumenes	
	flounder_paralichthys, vaccines, paralichthys_olivaceus, olive			
t_27 Antioxidant system	antioxidant, activities, levels, activity, carp, grass, grass_carp, increased, oxidative, glutathione, liver, mrna, decreased, superoxide, dismutase, superoxide_dismutase, expression, dietary, hepatic, effects	2.0 (27)	566 (28)	Health
t_28 Seabream culture	sea, mussel, fish, bream, mussels, sea_bream, farm, cages, cage, farms, species, mytilus, sites, gilthead_sea, gilthead, benthic, sediment, site, red, aurata	1.9 (28)	727 (25)	Specific aquatic organisms
t_29 Heritability	weight, genetic, selection, body, traits, breeding, families, body_weight, length, growth, age, family, strain, heritability, effects, variation, estimated, size, correlations, selected	1.9 (29)	767 (24)	Genetic
t_30 Flatfish culture	cells, cell, sole, halibut, observed, development, hippoglossus, number, microscopy, electron, solea, atlantic_halibut, histological, tissue, senegalese, senegalese_sole, morphology, morphological, deformities, stages	1.9 (30)	669 (27)	Specific aquatic organisms
t_31 Carpa	ponds, fish, carp, pond, production, common, common_carp, perch, silver, carpio, cyprinus, cyprinus_carpio, culture, yield, polyculture, carp_cyprinus, fingerlings, stocked, cobia, rice	1.7 (31)	495 (30)	Specific aquatic organisms
t_32 Amino acids requirement	amino, acid, amino_acid, amino_acids, acids, lysine, dietary, protein, methionine, requirement, phosphorus, taurine, arginine, diet, levels, red, free, drum, essential, met	1.5 (32)	435 (31)	Nutrition
t_33 Fish oil	fish, oil, muscle, fat, fish_oil, fillet, quality, liver, tissue, lipid, content, flesh, oils, fillets, vitamin, vegetable, tissues, composition, sensory, storage	1.4 (33)	361 (33)	Nutrition
t_34 Trout	trout, rainbow, rainbow_trout, oncorhynchus, mykiss, oncorhynchus_mykiss, trout_oncorhynchus, fish, brown, salmo, trout_salmo, walbaum, coho, salmon, brown_trout, salmon_oncorhynchus, chinook, mykiss_walbaum, coho_salmon, chinook_salmon	1.4 (34)	252 (35)	Specific aquatic organisms
t_35 Tilapia	tilapia, oreochromis, nile, niloticus, nile_tilapia, oreochromis_niloticus, prawn, tilapia_oreochromis, prawns, rosenbergii, macrobrachium, macrobrachium_rosenbergii, freshwater, prawn_macrobrachium, freshwater_prawn, springer, agalactiae, switzerland, giant, milkfish	1.3 (35)	106 (40)	Specific aquatic organisms

Tópicos	Top-20 terminos	Share (Ranking)		Temas
		Proporcion(%)	Número de resumenes	
t_36 Digestive enzymes	digestive, activity, enzyme, activities, enzymes, striped, bass, intestinal, intestine, protease, trypsin, amylase, striped_bass, tract, digestive_enzyme, lipase, digestive_enzymes, digestion, gland, enzyme_activities	1.2 (36)	297 (34)	Nutrition
t_37 European sea bass and seabream culture	abalone, bass, sea, sea_bass, haliotis, seabream, abalone_haliotis, european, labrax, dicentrarchus, dicentrarchus_labrax, gilthead, gilthead_seabream, bass_dicentrarchus, aurata, european_sea, discus, sparus, sparus_aurata, species	0.9 (37)	137 (38)	Specific aquatic organisms
t_38 Effect of light	light, crayfish, photoperiod, catfish, eels, clarias, intensity, anguilla, african, gariepinus, african_catfish, charr, arctic, clarias_gariepinus, light_intensity, catfish_clarias, continuous, clarkii, arctic_charr, dark	0.9 (38)	201 (36)	Husbandry protocols
t_39 Crabs	crab, crabs, astaxanthin, red, sinensis, chinese, colour, color, carotenoid, green, skin, black, blue, pigmentation, white, mitten, eriocheir, scylla, carotenoids, chinese_mitten	0.8 (39)	160 (37)	Specific aquatic organisms
t_40 Sea cucumber	sea, copper, japonicus, cu, cucumber, sea_cucumber, zn, iron, zinc, cucumbers, sea_cucumbers, carp, auratus, carassius, apostichopus, apostichopus_japonicus, carassius_auratus, goldfish, cucumber_apostichopus, bream	0.8 (40)	108 (39)	Specific aquatic organisms

Cada resumen de cada uno de los 38819 artículos está compuesto por los 40 tópicos descritos en la tabla 7 y la asignación de un artículo a un tópico determinado se hace en base al cuál es el tópico cuya proporcion sea la mayor en dicho artículo. A manera de ejemplo la Figura 7 muestra uno de los resúmenes compilados elegido al azar (el de Papatryphon *et al.*, (2000)) y las proporciones asignadas correspondientes de los 40 tópicos, así como los 30 términos más frecuentes. Encontramos que el tema 2 (“diet”) fue, con mucho, el tópico dominante para el resumen de Papatryphon *et al.*, (2000) seguido por el tópico 3 (“Growth performance“), el tópico 9 (“Diet formulation”) y el tópico 36 (“Digestive enzymes”). Dado que el tópico 2 fue el predominante (64.8%), consideramos que el artículo pertenecia a ese tópico.

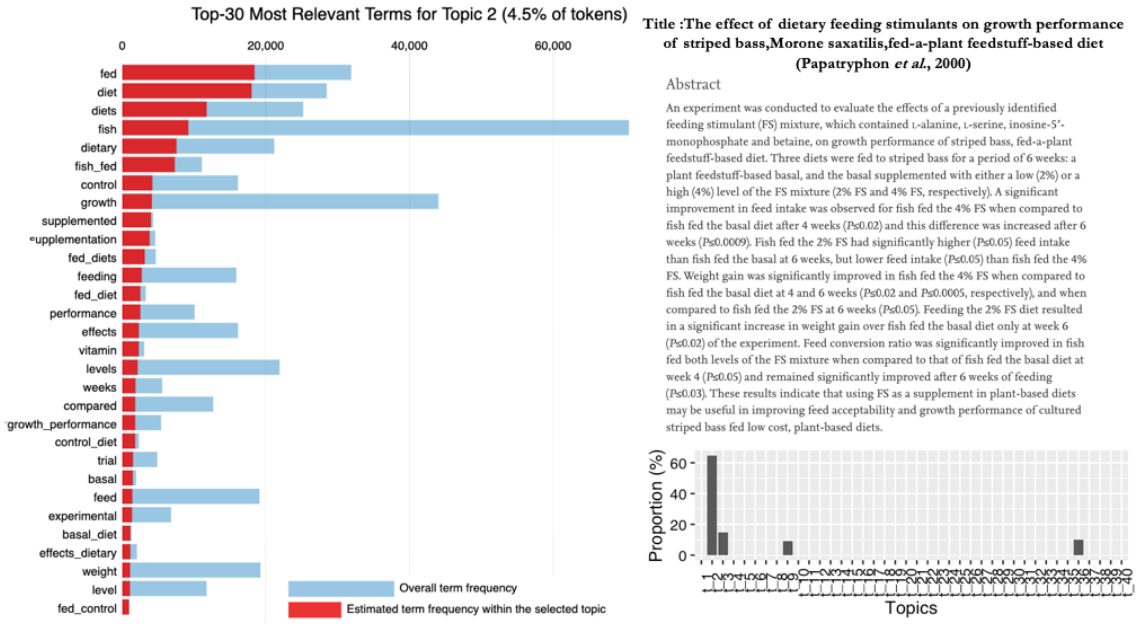

Figura 7. Ejemplo de la clasificación obtenida a partir de la evaluación de un resumen del corpus. Se muestra la proporción del tópico correspondiente (derecha abajo) y la frecuencia de terminos más frecuentes del tópico seleccionado (izquierda).

Los 40 tópicos encontrados se pueden agrupar en ocho temas generales: *Specific aquatic organisms* (n = 11), *Nutrition* (n = 8), *Husbandry protocols* (n = 6), *Health* (n = 6), *Genetics* (n = 4), *Fish performance* (n = 3), *Reproduction* (n = 1) y *Aquaculture production* (n = 1).

La Figura 8 muestra una representación bidimensional, en el plano 1-2, obtenida a partir del someter a la matriz de distancias entre tópicos a un *multidimensional scaling*. su visualización nos permite encontrar patrones de distribución de probabilidad similares sobre las palabras de los tópicos latentes. Consecuentemente, se encontró que se forman ciertos grupos, como por ejemplo el conformado por: [2 (“Diet”), 3 (“Growth performance”), 9 (“Diet formulation”), 13 (“Diet composition”), 24 (“Fatty acids”), 32 (“Amino acids requirement”), 33 (“Fish oil”)] al que podriamos denominar cluster de

crecimiento y nutrición. El tópico más aislado del resto es el 27 (“Antioxidant system”) que podría estar indicando el uso de palabras muy específicas dentro de ese tópico.

En el tópico “Specific aquatic organisms”, encontramos especies o grupos de especies tales como moluscos (bivalvos, camarones y cangrejos), peces (salmón, dorada, peces planos, carpa, trucha y tilapia) y también holoturoides (pepino de mar), que pueden tener alta probabilidad de ocurrencia porque representan especies de gran valor económico (Turpie *et al.*, 2003).

Natale *et al.* (2012) realizaron una revisión de la literatura en acuicultura utilizando métodos bibliométricos como el modelo de tema de análisis de co-citas y el LSA, encontrando, de forma similar a nuestro estudio, que la investigación en acuicultura está relacionada con los temas: genética, reproducción, crecimiento, fisiología, sistemas agrícolas y medio ambiente, nutrición, calidad del agua y salud. El número de tópicos informados por Natale *et al.* (2012) fue de seis. Sin embargo, los autores no mencionan cómo se obtuvo ese valor. También encontraron en su LSA las mismas especies que registramos en nuestros temas, con excepción del pepino de mar, el cangrejo, la dorada y *la platija*.

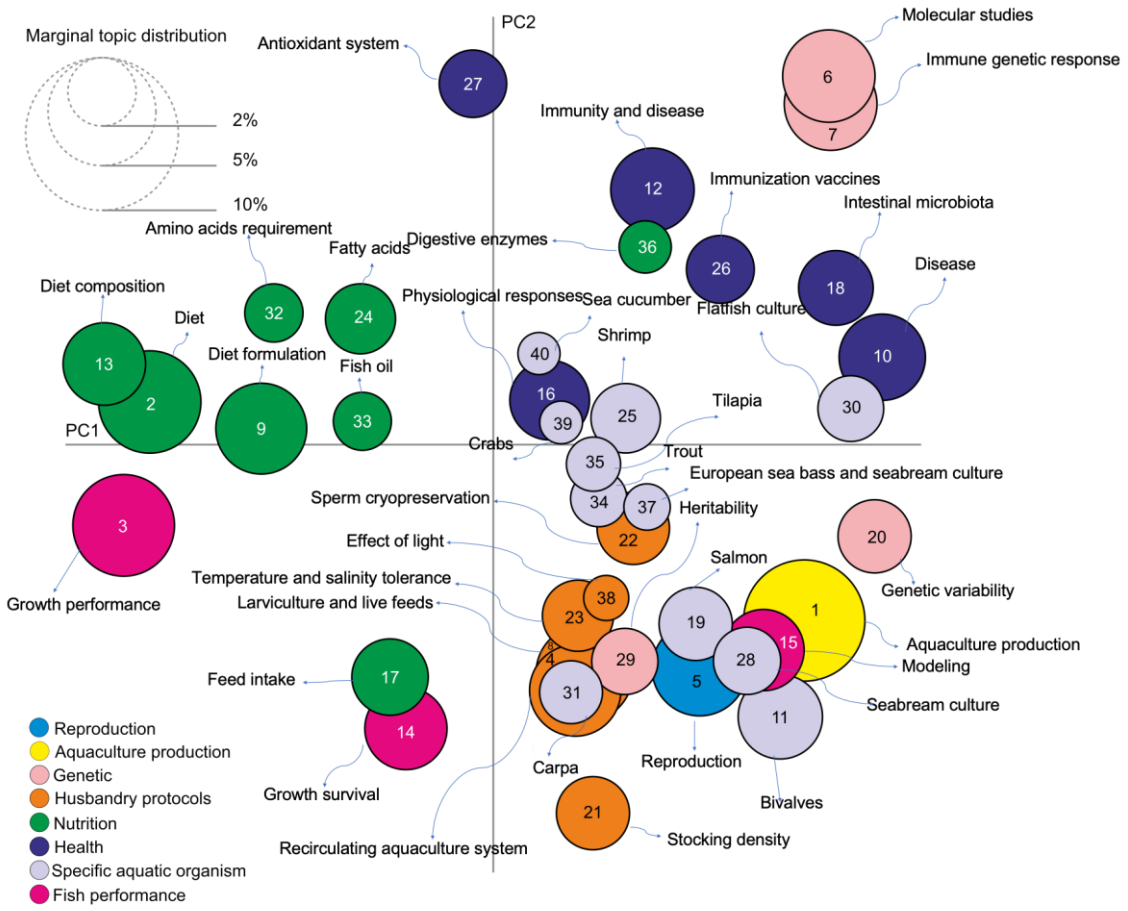

Figura 8. Representación bidimensional de los 40 tópicos via *multidimensional scaling*. La superficie de los círculos indica la prevalencia del tópico mientras que la distancia entre ellos representa la similitud.

5.2 TENDENCIAS TEMPORALES DE LOS TÓPICOS

Era importante comprender la tendencia general de la investigación en acuicultura, por lo que llevamos a cabo un estudio de la dinámica de los tópicos a lo largo del tiempo en términos de su proporción. El aumento de la proporción de un tópico indica que se trata de un campo de investigación emergente, mientras que su disminución muestra una tendencia de interés de investigación más baja. Encontramos que las probabilidades de 18 los tópicos han aumentado gradualmente con el tiempo, 14 mostraron tendencias

decrecientes y sólo 8 no presentaron tendencia (Figuras 9 y 10). Lo anterior podría proporcionar implicaciones convenientes para los investigadores,

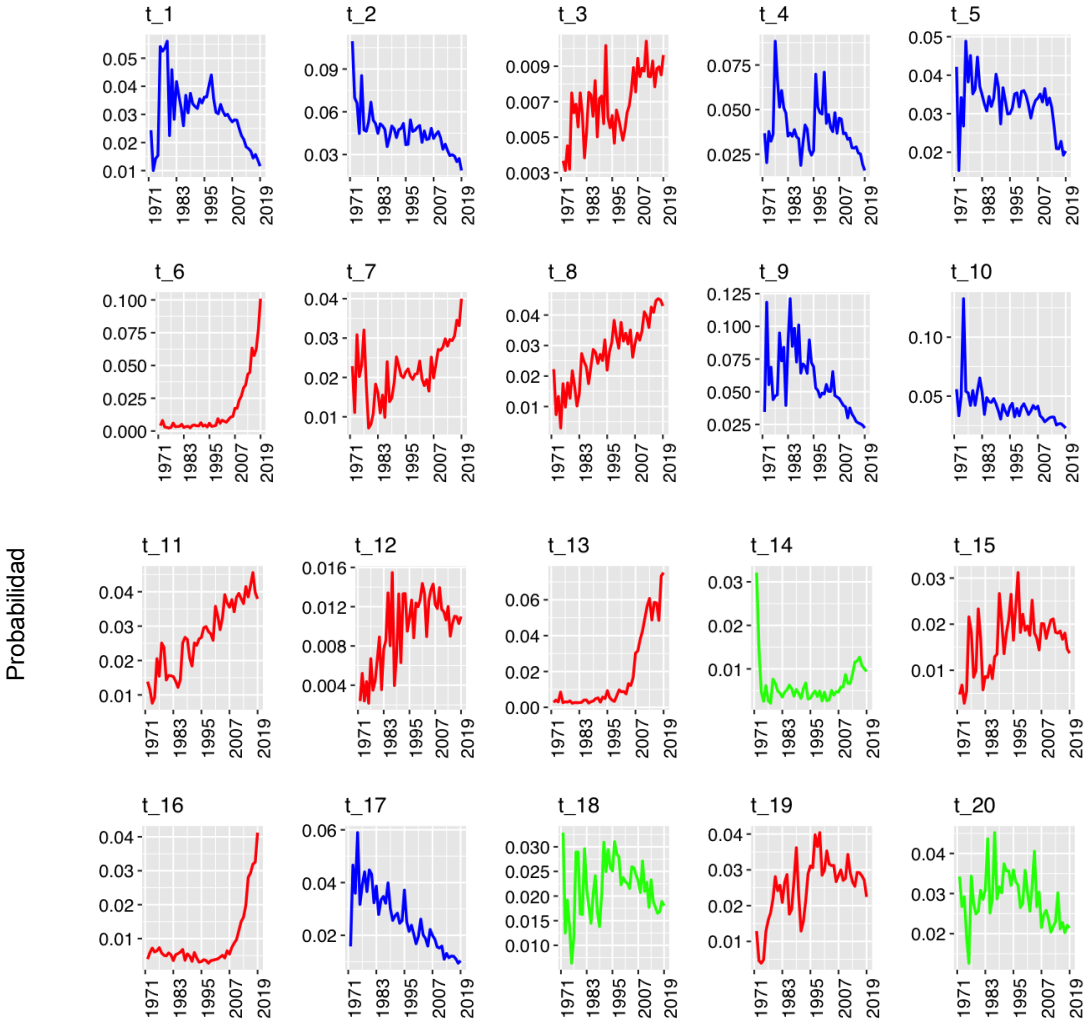

Figura 9. Tendencias de los tópicos (1-20) de investigación en acuicultura para 14 revistas especializadas en acuicultura en el período 1970-2019. Los colores de las líneas indican tendencia, rojo (creciente), azul (decreciente) y verde (fluctuante).

editores de revistas y formuladores de políticas en el campo acuícola dado que los investigadores pueden juzgar si sus investigaciones actuales se encuentran clasificadas como de tendencia creciente o decreciente, y seleccionar revistas apropiadas para enviar el resultado de sus trabajos. Dentro de los tópicos con tendencia decreciente

encontramos que los tópicos (1), (2), (17), (22), (25), (26), (30) y (34), inician con un patrón de alta frecuencia que fue gradualmente descendiendo, lo que indica una posible disminución en su popularidad con el transcurso del tiempo dentro del comunidad cientifica.

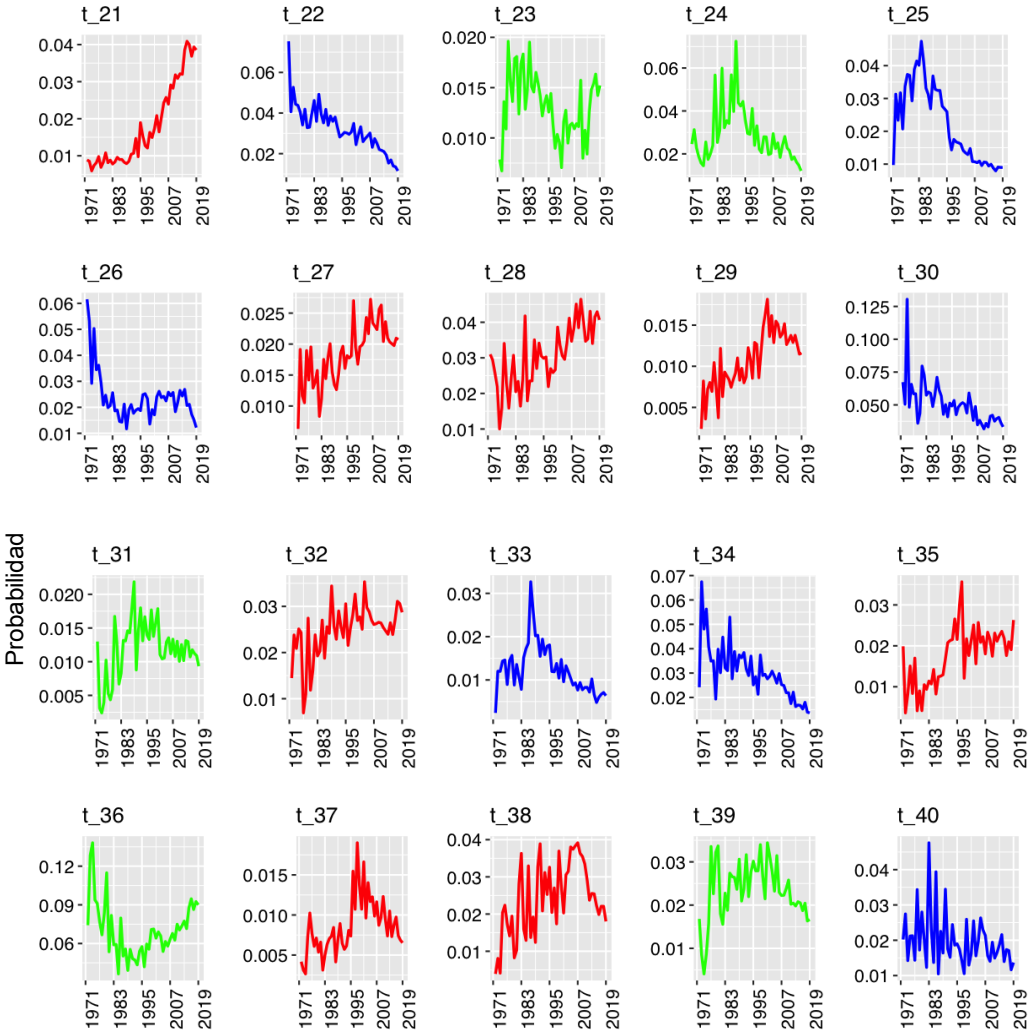

Figura 10. Tendencias de los tópicos (21-40) de investigación en acuicultura para 14 revistas especializadas en acuicultura en el período 1970-2019. Los colores de las líneas indican tendencia, rojo (creciente), azul (decreciente) y verde (fluctuante).

Es complejo establecer la popularidad de un tema si solo consideramos su tendencia de forma aislada, como es el caso de los tópicos 9 (“Diet formulation”) y 30 (“Flatfish

culture"), que mostrarón una tendencia decreciente, sin embargo, tienen una probabilidad relativamente grande. Teniendo esto en cuenta, medimos la popularidad, que considera conjuntamente la tendencia y la probabilidad del tópico.

Los tópicos más populares fueron: 36 ("Digestive enzymes"), 13 ("Diet composition"), 28 ("Seabream culture"), 8 ("Larviculture and live feeds") y el 6 ("Molecular studies"), mientras que los que tienen la menor la popularidad fueron el tópico 34 ("Trout"), 26 ("Immunization vaccines"), 17 ("Feed intake"), 25 ("Shrimp") y el 33 ("Fish oil") (Tabla 4). Observamos que, en general, todos los tópicos con tendencias crecientes se encuentran en las primeras clasificaciones de popularidad, independientemente de su promedio de probabilidad, con la excepción del tópico 1 ("Aquaculture production") que tiene la probabilidad más alta y su tendencia fue fluctuante (Tabla 8)

Curiosamente, encontramos que los tópicos 9 ("Diet formulation"), 30 ("Flatfish culture") y 2 ("Diet"), mostraron las probabilidades promedio segunda, tercera y cuarta más altas se clasificaron en las posiciones 25º, 26º y 29º de popularidad.

.

Tabla 8. Popularidad de los 40 Tópicos

Rank	Topic	Average probability	Normalized probability	Trend score	Popularity	Rank	Topic	Average probability	Normalized probability	Trend score	Popularity
1	t_36 Digestive enzymes	0,0722	1,000	0,67	1,670	21	t_20 Genetic variability	0,0265	0,366	0,67	1,036
2	t_13 Diet composition	0,0358	0,496	1	1,496	22	t_24 Fatty acids	0,0245	0,339	0,67	1,009
3	t_28 Seabream culture	0,0357	0,495	1	1,495	23	t_39 Crabs	0,0227	0,314	0,67	0,984
4	t_8 Larviculture and live feeds	0,0354	0,490	1	1,490	24	t_18 Intestinal_microbiota	0,0214	0,296	0,67	0,966
5	t_6 Molecular studies	0,0347	0,481	1	1,481	25	t_9 Diet formulation	0,0433	0,599	0,33	0,929
6	t_11 Bivalves	0,0345	0,478	1	1,478	26	t_30 Flatfish culture	0,0428	0,593	0,33	0,923
7	t_19 Salmon	0,0280	0,387	1	1,387	27	t_40 Sea cucumber	0,0179	0,248	0,67	0,918
8	t_32 Amino acids requirement	0,0270	0,375	1	1,375	28	t_23 Temperature and salinity tolerance	0,0129	0,178	0,67	0,848
9	t_21 Stocking density	0,0270	0,374	1	1,374	29	t_2 Diet	0,0374	0,518	0,33	0,848
10	t_38 Effect of light	0,0265	0,367	1	1,367	30	t_31 Carpa	0,0121	0,168	0,67	0,838
11	t_7 Immune genetic response	0,0262	0,363	1	1,363	31	t_10 Disease	0,0335	0,464	0,33	0,794
12	t_27 Antioxidant system	0,0208	0,288	1	1,288	32	t_14 Growth survival	0,0072	0,100	0,67	0,770
13	t_35 Tilapia	0,0208	0,288	1	1,288	33	t_5 Reproduction	0,0294	0,407	0,33	0,737
14	t_15 Modeling	0,0179	0,248	1	1,248	34	t_1 Aquaculture production	0,0247	0,342	0,33	0,672
15	t_16 Physiological responses	0,0157	0,217	1	1,217	35	t_22 Sperm cryopreservation	0,0242	0,335	0,33	0,665
16	t_29 Heritability	0,0126	0,174	1	1,174	36	t_34 Trout	0,0233	0,323	0,33	0,653
17	t_12 Immunity and disease	0,0111	0,154	1	1,154	37	t_26 Immunization vaccines	0,0207	0,287	0,33	0,617
18	t_4 Recirculating aquaculture system	0,0341	0,472	0,67	1,142	38	t_17 Feed intake	0,0180	0,249	0,33	0,579
19	t_37 European sea bass and seabream culture	0,0093	0,129	1	1,129	39	t_25 Shrimp	0,0145	0,201	0,33	0,531
20	t_3 Growth performance	0,0080	0,111	1	1,111	40	t_33 Fish oil	0,0099	0,137	0,33	0,467

5.3 TENDENCIAS TEMPORALES DE LOS TÓPICOS DENTRO DE LAS REVISTAS

En el mapa de calor presentado en la Figura 11, se muestra la distribución de tópicos por revista, en donde el color del píxel representa la probabilidad (el verde representa el más pequeño y el rojo el más grande) de que un tópico determinado se mencione en una revista en particular. Algunas revistas como *Aquaculture*, *Aquaculture Research*, *Aquaculture International*, *Aquaculture Reports*, *Aquaculture Nutrition* y *Journal of the World Aquaculture Society,* tienen un alcance relativamente amplio, mientras que otras se centran en tópicos específicos como *Aquaculture Economics and Management*, *Reviews in Aquaculture and Fisheries Science*, y *Reviews in Aquaculture*, que se centran específicamente en el tópico 1 (“Aquaculture production”). Del mismo modo, *Aquaculture Environment Interactions* y *Aquacultural Engineering*, se concentran en los tópicos 4 (“Recirculating aquaculture system”) y 15 (“Modelling”).

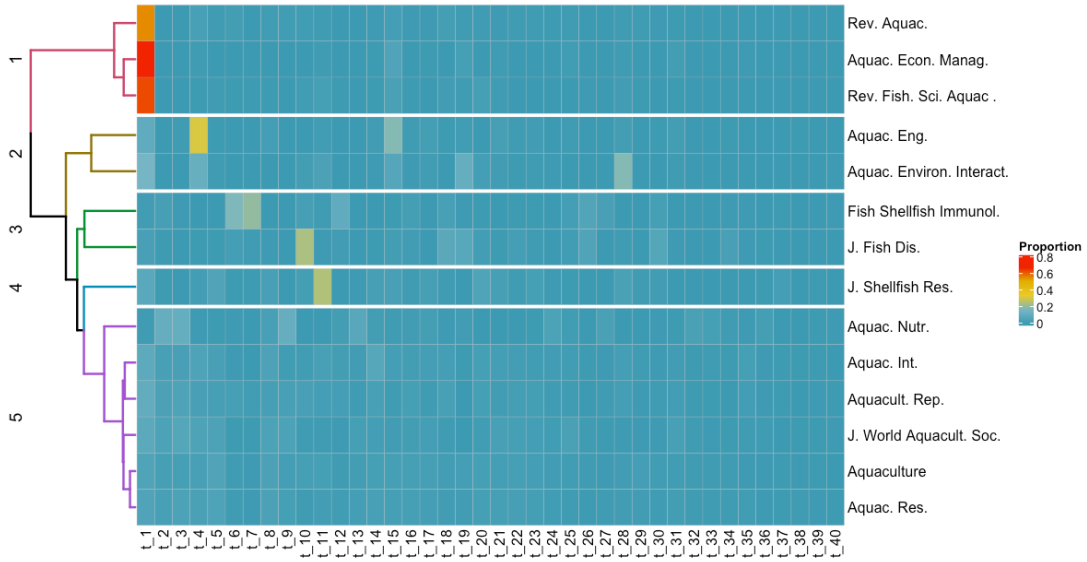

Figura 11. Mapa de calor de la distribución y dendograma de similitud de tópicos para 14 revistas especializadas en acuicultura en el período 1972-2019

El dendograma resultante del análisis de cluster se muestra en la parte izquierda de la Figura 11. Como puede observarse, las revistas fueron clasificadas en 5 grupos siendo la revista más singular entre las 14 consideradas el *Journal of Shellfish Research* (cluster 4), Las revistas restantes se pueden clasificar en cuatro grupos. Por ejemplo, podemos identificar, el grupo 2 (*Aquaculture Environment Interactions–Aquacultural Engineering*) y 3 (*Fish and Shellfish Immunology–Journal of Fish Diseases*) y el grupo 5, con el mayor número de revistas, es decir, *Aquaculture Internationa*l, *Journal of the World Aquaculture Society*, *Aquaculture Nutrition*, *Aquaculture Research*, *Aquaculture Reports* y *Aquaculture*.

La dinámica temporal de los tópicos en cada revista, se presenta en la Figura 12 mediante la representación gráfica de áreas apiladas que representan los 40 tópicos, en orden descendente. Observamos que la aparición de la revista *Fish and Shellfish Immunology* es el principal impulsor de las variaciones de proporción de tópicos, de los tópicos 6 ("Molecular studies") y 7 ("Immune genetic response"). También encontramos que en algunas revistas la distribución de tópicos es relativamente uniforme (esto es, *Aquaculture* y *Journal of the World Aquaculture Society*). Otros tópicos han disminuido con el tiempo, por ejemplo, el tópico 14 (‘‘Growth survival") en *Aquaculture International* y el tópico 1 en *Aquaculture Research*.

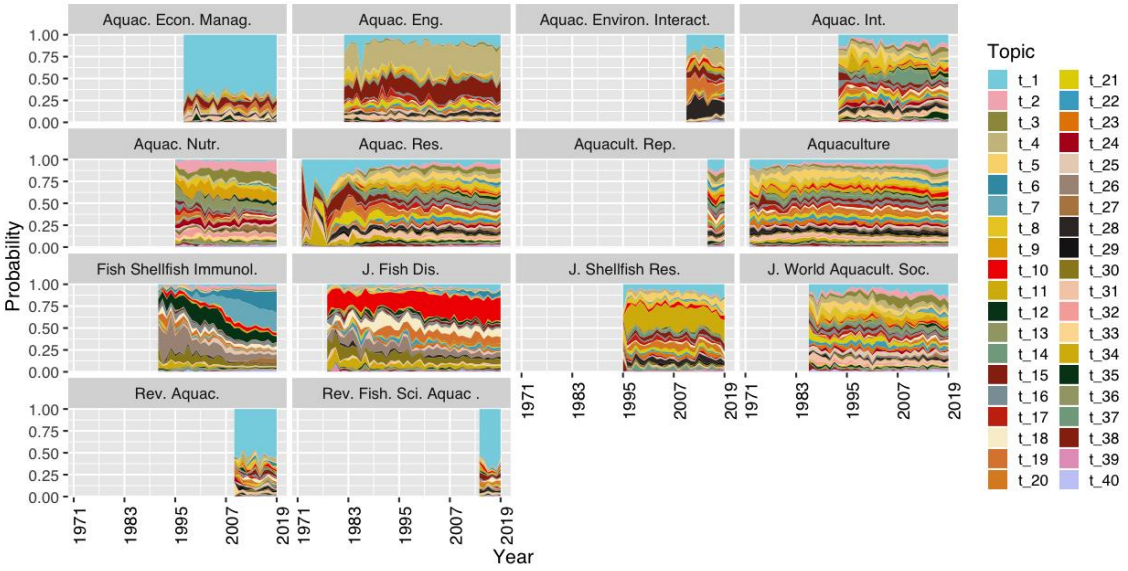

Figura 12. Distribución de tópicos para 14 revistas especializadas en acuicultura en el período 1972-2019.

Al comparar las revistas a nivel cuantitativo, en términos de entropía de la información, *Aquaculture* y *Aquaculture Research* muestran el mayor valor, lo que significa que cubren una amplia variedad de tópicos, mientras que *Aquaculture Economics and Management* y *Reviews in Fisheries Science and Aquaculture* presentan los valores más bajos de entropía. En el último caso, esto podría explicarse por el hecho de que los tópicos en esas revistas están principalmente relacionados más con las ciencias pesqueras que con la acuicultura (Figura 13).

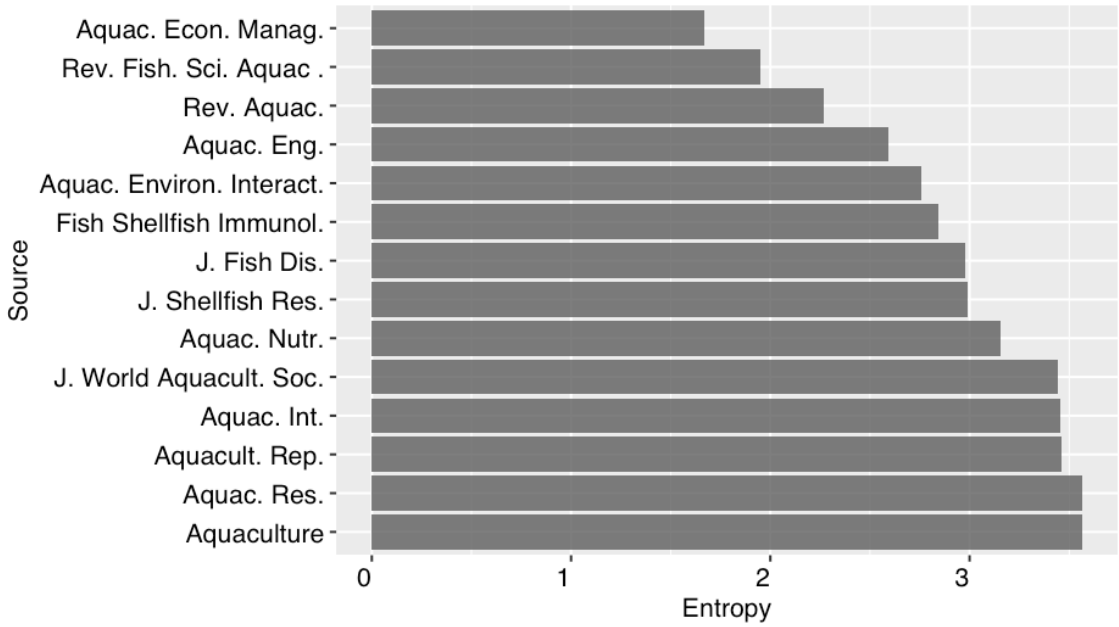

Figura 13. Entropía de información para 14 revistas especializadas en acuicultura en el período 1972-2019

5.4 TENDENCIAS TEMPORALES DE LOS TEMAS DENTRO DE LAS REVISTAS

Para analizar el comportamiento de los temas a lo largo de la serie temporal estudiada, se trabajó con 10 de las 14 revistas puesto que 4 de ellas estaban fuera del rango temporal de tiempo escogido. Realizamos en primer lugar el análisis estático, (primera fase de un Biplot dinámico), utilizando como factorización el análisis HJ-Biplot, fijando como situación de referencia la correspondiente al quinquenio Q4 (2014-2019) que es la situación más reciente que tenemos.

El plano 1-2 recoge una inercia del 77.92 % de la que el 59.51 % se corresponde al primer eje factorial (Figura 14).

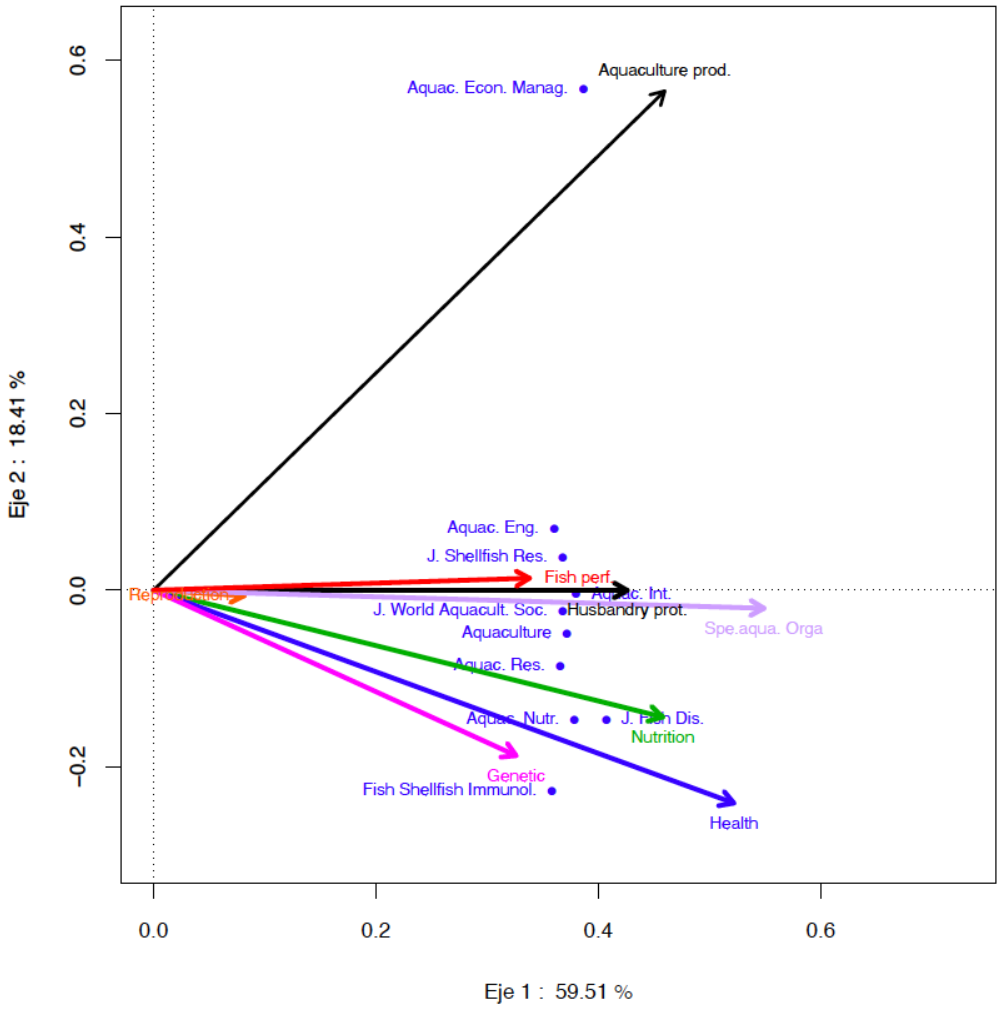

Figura 14. Etapa 1: Análisis estático de los temas con referencia al quinquenio 2014-2019.

En la Tabla 9 se muestra la inercia de los temas en los cuatro primeros ejes, donde se han resaltado aquellas que acumulan más de 500 en el plano 1-2. Prácticamente todas las variables bien representadas son de eje 1 salvo *Aquaculture production* que caracteriza el eje 2. Esta variable es una variable de plano y tiene una inercia acumulada de 965, mientras que la variable con menor representación en el plano es *Reproduction* con una inercia acumulada de 501.

Tabla 9. Inercia de las variables en los cuatro primeros ejes.

	Eje			
Tema	1	2	3	4
Aquaculture prod.	384	**581**	22	10
Reproduction	498	3	11	22
Nutrition	**566**	56	177	199
Health	**691**	147	101	0
Genetic	436	144	264	1
Husbandry prot.	**624**	0	142	210
Spe.aqua. Orga	**868**	1	6	19
Fish perf.	**765**	1	188	19

Sobre el gráfico biplot obtenido en el primer paso proyectamos cad uno de los temas en cada uno de los quinquenios, obteniendo así sus posiciones y sus trayectorias (Figura 15).

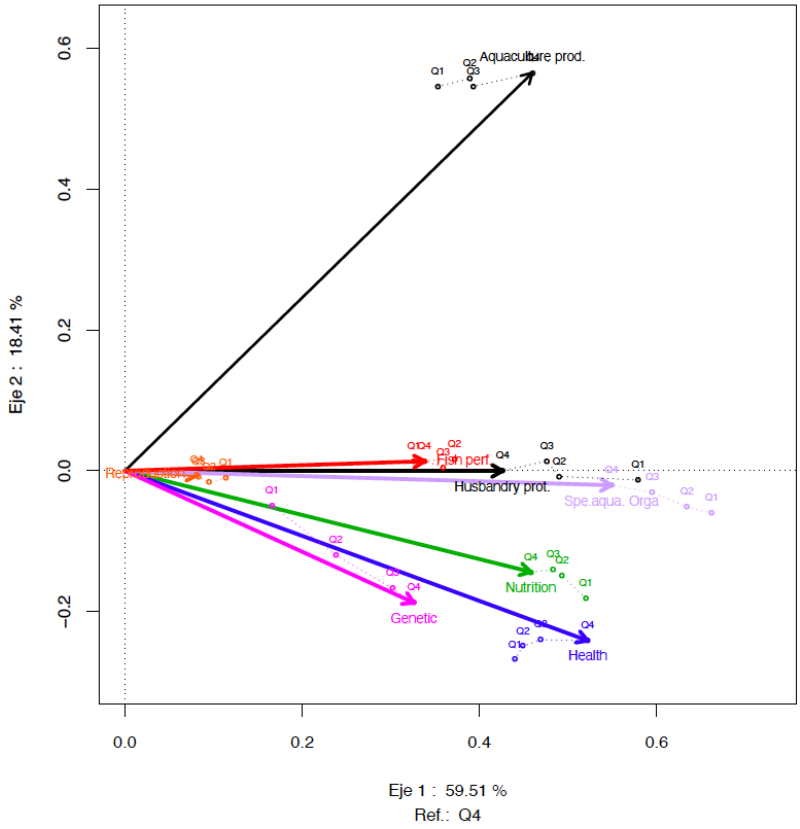

Figura 15. Etapa 2: Análisis dinámico, vista general de las trayectorias de los temas en todos los quinquenios, tomando como referencia el quinquenio Q4 (2014-2019). Q1 es el primer quinquenio 2000-2004, Q2 (2005-2009) y Q3 (2010-2014)

En la Tabla 10 podemos ver los coeficientes de determinación de las regresiones realizadas para calcular los marcadores de las proyecciones de las variables (temas), que nos sirven como medida de la calidad de representación, además se especifican, entre paréntesis, los p-valores de los ANOVAs de dichas regresiones, resaltando en rojo aquéllas que fueron significativas al 0.05.

Los valores de los coeficientes de determinación resultantes han sido bajos (menores del 50%) y no significativos para las variables *Aquaculture production*, *Reproduction* y *Nutrition* en todos los quinquenios, mientras que para los temas restantes todos los coefecientes de determinación (exepto *Genetic* 2004-2009) fueron significativos.

Tabla 10. Coeficientes de determinación de las regresiones para las variables (temas), en el plano 1-2. Entre parentesis los p-valores de los ANOVAs de las regresiones, resaltando en rojo aquellas que fueron significativas.

	Quinquenios			
Temas	2000-2004	2005-2009	2010-2014	2015-2019
Aquaculture prod.	0,4644 (0,1125)	0,4597 (0,116)	0,4525 (0,1214)	0,3856 (0,1818)
Reproduction	0,3016 (0,2846)	0,2517 (0,3625)	0,3671 (0,2016)	0,5373 (0,0674)
Nutrition	0,0684 (0,7804)	0,0825 (0,7399)	0,0931 (0,7103)	0,0871 (0,727)
Health	0,7446 (0,0084)	0,7181 (0,0119)	0,6716 (0,0203)	0,6255 (0,0321)
Genetic	0,5579 (0,0574)	0,5923 (0,0433)	0,5817 (0,0473)	0,6115 (0,0365)
Husbandry prot.	0,6481 (0,0258)	0,71 (0,0131)	0,6896 (0,0167)	0,6377 (0,0286)
Spe.aqua. Orga	0,8258 (0,0022)	0,7811 (0,0049)	0,741 (0,0088)	0,7189 (0,0118)
Fish perf.	0,9049 (0,0003)	0,8133 (0,0028)	0,7582 (0,0069)	0,8898 (0,0004)

En la figura 16 se presentan las trayectoria de todos los temas de forma ampliada para dar más claridad. Analizamos solo las trayectorias de los temas suficientemente bien representados en el gráfico biplot.

En la Figura 16(a) observamos la evolución de los temas *Health* y *Genetic*, en ambos se aprecia la evolución positiva (aumento del interés) que han tenido estos temas a lo largo del tiempo. En la figura 16(b) los temas *Husbandry protocols* y *Specific aquatics organism* muestran tendencia a la disminución (ha disminuido su interés), mientras que *Fish performance* tiene una trayectoria bastante corta, incrementándose al principio y disminuyendo posteriormente para acercarse a los valores del primer quinquenio.

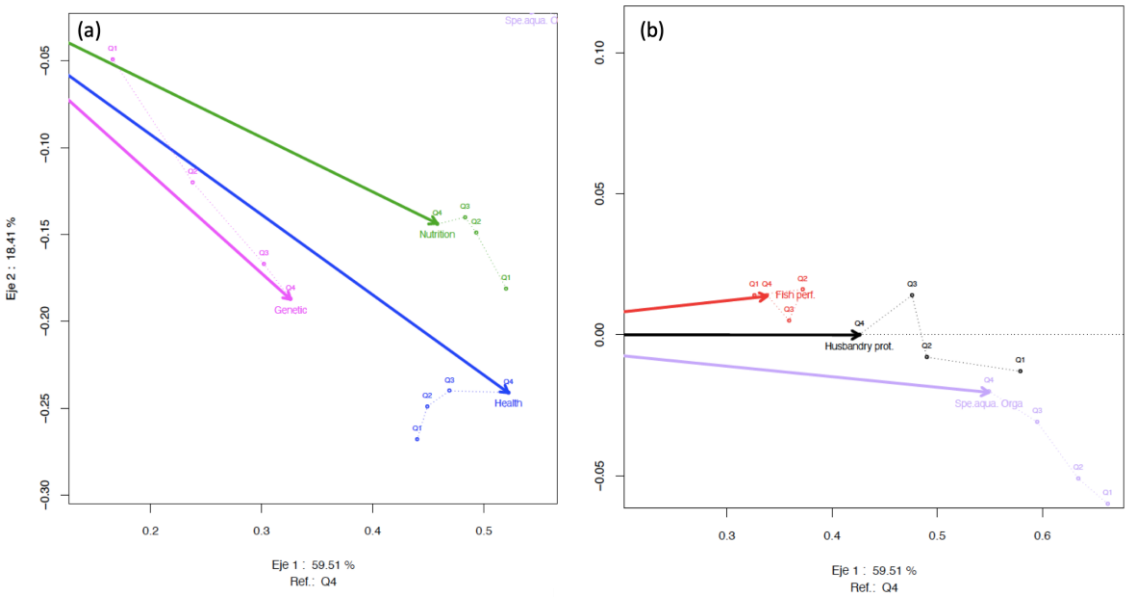

Figura 16. Análisis dinámico, vista general ampliada de la trayectoria de los temas. Q1 (2000-2004), Q2 (2005-2009) , Q3 (2010-2014) y Q4 (2015-2019)

6 CONCLUSIONES

1. En este trabajo se ha puesto de manifiesto la utilidad del LDA para la revisión exploratoria de literatura científica en acuicultura, dado que crea tópicos coherentes que resumen la colección de documentos evaluados, y disminuye drásticamente el tiempo utilizado para la revisión de los artículos.

2. Este trabajo permitió identificar y analizar 40 tópicos latentes y sus tendencias en el campo de la investigación en acuicultura, 18 tópicos aumentaron su interés con el paso del tiempo, 14 disminuyeron en interés y 8 no presentaron tendencia

3. El análisis MDS permitió validar el etiquetado previo de los tópicos, al mostrar agrupaciones coherentes y superposición de los nodos, lo que indica distribuciones de palabras similares.

4. Mediante el análisis de clusters, pudimos identificar 5 grupos de revistas : grupo 1 [*Review Aquaculture-Aquaculture Economic management-Review Fisheries Sciences and Aquaculture*][*Journal of Shellfish Research]* ; *grupo 2 [Aquaculture Environment Interactions–Aquacultural Engineering]* ; grupo 3 [*Fish and Shellfish Immunology–Journal of Fish Diseases*]; grupo 4 [*Journal Shellfish Research*] y el grupo con el mayor número estuvo conformado por las revistas [*Aquaculture International-Journal of the World Aquaculture Society-Aquaculture Nutrition- Aquaculture Research- Aquaculture Reports –Aquaculture].*

5. Los hallazgos encontrados en cuanto a las tendencias podría ser de gran utilidad para los editores de las revistas, ya que podrían confirmar si sus publicaciones coinciden con su política editorial en lo que referente a los tópicos o bien podría ayudarles a dar nueva perspectiva a la editorial.

6. Análogamente, los investigadores en el campo de la acuicultura, pueden juzgar si sus publicaciones actuales se encuentran clasificadas como de tendencia interés creciente o decreciente y seleccionar aquéllas revistas más apropiadas para enviar el resultado de sus trabajos.

7. Los tópicos identificados se pueden agrupar en temas más generales que corresponden a un esquema de clasificación del sector de la acuicultura, mostrando que la investigación fundamentalmente está enfocada en la dimensión biológica (Specific aquatic organisms, Nutrition, Husbandry protocols, Health and Genetics).

8. Se ha encontrado que la aparición de la revista *Fish and Shellfish Immunology* en 1991 marcó un hito en el interés de los tópicos 6 ("Molecular studies") y 7 ("Immune genetic response"), proporcionándoles un gran impulso.

9. Los tópicos más populares fueron "Digestive enzymes", "Diet composition", "Seabream culture", "Larviculture and live feeds" y "Molecular studies" .

10. Todos los tópicos con interés crecientese encuentran en las primeras clasificaciones de popularidad, independientemente de su promedio de probabilidad, con la excepción del tópico 1 “Aquaculture production” que tiene la probabilidad más alta y su tendencia fue fluctuante.

11. El biplot dinámico permitió conocer no sólo cuáles eran los temas predominantes en las diferentes revistas en el último quinquenio (2014-2019), sino también estudiar la evolución del interés por los diferentes temas. Se ha encontrado que los temas relacionados con genética (Genetic) y salud (Health) y protocolos de crianza (Husbandry protocols) han aumentado su interés a lo largo de la serie temporal, mientras que ha disminuido en aquéllos relacionados con nutrición (Nutrition), protocolos de crianza (Husbandry protocols) y organismos acuáticos específicos (Specifics aquatic organism). El rendimiento de peces (Fish performance) mostro un comportamiento fluctuante.

REFERENCIAS

Adams, J., Khan, H. T., Raeside, R., & White, D. I. (2007). *Research methods for graduate business and social science students*. SAGE publications India.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, *7*(1), 1-16.

Anthes, G. (2010). Topic models vs. unstructured data. *Communications of the ACM*, *53*(12), 16-18.

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2012). On Smoothing and Inference for Topic Models. In *UAI '09 Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 27–34). Montreal, Quebec, Canada: AUAI Press Arlington

Bar-Ilan, J. (2008). Which h-index? -A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74 (2), 257-271. https://doi.org/10.1007/s11192-008-0216-y

Bergamaschi, S., & Po, L. (2014, April). Comparing LDA and LSA topic models for content-based movie recommendation systems. In *International conference on web information systems and technologies* (pp. 247-263). Springer, Cham.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Muller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. J. Open Source Software, 3 (30), 774

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, *37*(4), 573-595

Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: applications and theory*. John Wiley & Sons.

Blei. DM. Ng. OH. & Jordan. MI (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3 (Jan). 993-1022.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, *1*(1), 121-143.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17-35.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In *Text mining* (pp. 101-124). Chapman and Hall/CRC.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Board, NS (2012). Science and Engineering Indicators 2012. NS Foundation (Ed.). Arlington, VA: National Science Foundation

Brants, T. (2005). Test data likelihood for PLSA models. *Information Retrieval*, *8*(2), 181-196

Buntine, W. (2009, November). Estimating likelihoods for topic models. In *Asian Conference on Machine Learning* (pp. 51-64). Springer, Berlin, Heidelberg.

Chang, J., & Blei, D. (2009, April). Relational topic models for document networks. In *Artificial Intelligence and Statistics* (pp. 81-88).

Chemudugunta, C. (2010). *Text mining with probabilistic topic models: applications in information retrieval and concept modeling*. Lambert Academic Publishing.

Chen, L. C. (2017). An effective LDA-based time topic model to improve blog search performance. *Information Processing & Management*, *53*(6), 1299-1319.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the american statistical association*, *90*(432), 1313-1321.

Chien, J. T., & Chueh, C. H. (2008, December). Latent Dirichlet language model for speech recognition. In *2008 IEEE Spoken Language Technology Workshop* (pp. 201-204). IEEE.

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and trust. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems* - CHI'12 (pp 443-452.). Austin, TX: ACM Press

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(3), 411-436.

de Wildt, T. E., Chappin, E. J., van de Kaa, G., & Herder, P. M. (2018). A comprehensive approach to reviewing latent topics addressed by literature across multiple disciplines. *Applied Energy*, 228, 2111-2128.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, *41*(6), 570-606.

Jaime Egido (2017). dynBiplotGUI: Full Interactive GUI for Dynamic Biplot in R. *R package version 1.1.5*. https://CRAN.R-project.org/package=dynBiplotGUI

Egido, J., & Galindo, P. (2015). Dynamic Biplot. Evolution of the Economic Freedom in the European Union. *Current Journal of Applied Science and Technology*, 1-13.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems*, *21*(1), 70-86.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3), 453-467.

Galindo, M. P. (1986). Una alternativa de representacion simultanea: HJ-Biplot. *Qüestiió: quaderns d'estadística i investigació operativa*, 13-23.

Garfield, E. (1955). Citation indexes for science. *Science*, *122*(3159), 108-111.

Gaussier, E., & Yvon, F. (Eds.). (2013). *Textual information access: statistical models*. John Wiley & Sons.

Geman. S .. and Geman. D. (1984). Stochastic Relaxation. Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 6. 721-741

Girolami, M., & Kabán, A. (2003, July). On an equivalence between PLSI and LDA. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 433-434).

Griffiths. T. L.. & Steyvers. M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*. *101*(suppl 1). 5228-5235.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*(1), 1-35.

Harzing, AW, & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106 (2), 787-804.

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57).

Iwata, T., Saito, K., Ueda, N., Stromsten, S., Griffiths, T. L., & Tenenbaum, J. B. (2005). Parametric embedding for class visualization. In *Advances in neural information processing systems* (pp. 617-624).

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89-106.

Jones, T. (2019). textmineR: Functions for Text Mining and Topic Modeling. *R package version* 3.0.4. https://CRAN.R-project.org/package=textmineR

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Jurasky, D., & Martin, J. H. (2000). Speech and Language Processing: An introduction to natural language Processing. *Computational Linguistics and Speech Recognition. Prentice Hall, New Jersey*.

Kao, A., & Poteet, S. R. (Eds.). (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Lau, JH, Grieser, K., Newman, D., & Baldwin, T. (2011, June). Automatic labeling of topic models. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:* Human Language Technologies-Volume 1 (1536-1545 pp.). Association for Computational Linguistics.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

Lewis, SC, Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and Manual methods. *Journal of Broadcasting & Electronic Media*, 57 (1), 34-52

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and Its current applications in bioinformatics. SpringerPlus, 5 (1), 1608.

Li, W., & McCallum, A. (2006, June). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning* (pp. 577-584).

Lochbaum, K. E., & Streeter, L. A. (1989). Comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval. *Information Processing & Management*, *25*(6), 665-676

Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, *14*(2), 178-203.

Mather, ME, Parrish, DL, & Dettmers, JM (2008). Mapping the changing landscape of fish-related journals: Setting a course for successful communication of scientific information. Fisheries, 33 (9), 444-453.

Natale, F., Fiore, G., & Hofherr, J. (2012). Mapping the research on aquaculture. A Bibliometric analysis of aquaculture literature. Scientometrics, 90 (3), 983-999. https://doi.org/10.1007/s11192-011-0562-z

Newman, D., Smyth, P., Welling, M., & Asuncion, A. U. (2008). Distributed inference for latent dirichlet allocation. In *Advances in neural information processing systems* (pp. 1081-1088).

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(1), 3-26.

Papatryphon, E., & Soares Jr, J. H. (2000). The effect of dietary feeding stimulants on growth performance of striped bass, Morone saxatilis, fed-a-plant feedstuff-based diet. *Aquaculture*, *185*(3-4), 329-338.

Paré, G., Trudel, M. C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183-199. https://doi.org/10.1016/j.im.2014.08.008

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008, August). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 569-577).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137.

Salton, G. (1975). A vector space model for information retrieval. *Journal of the ASIS*, 613-620.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *In Proceedings of the Workshop on Interactive Language Learning,*

Shin, S. H., Kwon, O. K., Ruan, X., Chhetri, P., Lee, P. T. W., & Shahparvari, S. (2018). Analyzing sustainability literature in maritime studies with text mining. *Sustainability*, *10*(10), 3522.

Teh, Y. W., Newman, D., & Welling, M. (2007). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in neural information processing systems* (pp. 1353-1360).

Titov, I., & McDonald, R. (2008, April). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web* (pp. 111-120).

Torkkola, K. (2002, August). Discriminative features for document classification. In *Object recognition supported by user interaction for service robots* (Vol. 1, pp. 472-475). IEEE.

Turpie, J. K., Heydenrych, B. J., & Lamberth, S. J. (2003). Economic value of terrestrial and marine biodiversity in the Cape Floristic Region: implications for defining effective and socially optimal conservation strategies. *Biological conservation*, *112*(1-2), 233-251

Vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009, June). Reconstructing the giant: on the importance of rigour in documenting the literature search process. In *Ecis* (Vol. 9, pp. 2206-2217).

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009, June). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105-1112).

Wang, C., & Blei, D. M. (2011, August). Collaborative topic modeling for recommending scientific articles. *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 448-456).

Wang, C., Paisley, J., & Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, in PMLR*.

Wang, Y., Sabzmeydani, P., & Mori, G. (2007, October). Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *Workshop on Human Motion* (pp. 240-254). Springer, Berlin, Heidelberg.

Wei, X., & Croft, W. B. (2006, August). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 178-185).

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Xiong, H., Cheng, Y., Zhao, W., and Liu, J. (2019). Analyzing scientific research topics in manufacturing field using a topic model. *Computers & Industrial Engineering*.