



*Escuela Técnica Superior de Ingenieros de Caminos,
Canales y Puertos.*
UNIVERSIDAD DE CANTABRIA



APLICACIÓN DE LA LEY DE BENFORD A VARIABLES DEL ÁMBITO COSTERO.

Trabajo realizado por:

Laura Puente Mella

Dirigido:

Raúl Medina Santamaria

Jara Martínez Sánchez

Titulación:

**Máster Universitario
en Costas y Puertos**

Santander, Septiembre de 2022

TRABAJO FIN DE MASTER



*Escuela Técnica Superior de Ingenieros de Caminos,
Canales y Puertos.*
UNIVERSIDAD DE CANTABRIA



ÍNDICE

RESUMEN	3
ABSTRACT	4
1. INTRODUCCIÓN.....	5
2. LEY DE BENFORD.....	6
3. BASE DE DATOS.....	8
3.1. Fuente de información	8
3.2. Clasificación de las variables.....	9
3.3. Descripción de los datos.....	9
4. ANÁLISIS DE LOS DATOS.....	13
4.1. Metodología de análisis	13
4.2. Resultados del análisis	14
5. DISCUSIÓN	23
6. CONCLUSIÓN	25
REFERENCIAS.....	27

RESUMEN

Un patrón común en la distribución de frecuencias del primer dígito de distintos conjuntos de números reales, abre la posibilidad de demostrar la consistencia de una base de datos a través de dicha distribución. Este patrón, definido por la Ley de Benford, considera más probables los dígitos 1, 2 y 3, con un descenso de la frecuencia de aparición hasta el 9. En este estudio, se somete una base de datos con variables tanto físicas como socioeconómicas del ámbito costero y se verifica la conformidad de aquellas variables que cumplen las condiciones para la aplicación de la Ley de Benford. Esta base de datos proviene del Plan Estratégico Nacional para la Protección de la Costa Española considerando los Efectos del Cambio Climático y cuenta con información del litoral español en su totalidad. De ella se extraen variables de interés para la gestión costera referidas a playas, cuencas hidrográficas, clima marítimo, yacimientos submarinos, zonas potencialmente inundables, población y amarres en puertos deportivos. Estas variables se someten a pruebas de bondad de ajuste que determinan la conformidad con la Ley de Benford. Los resultados obtenidos muestran una conformidad alta para la longitud de playas, tramos de erosión en playas, área de las cuencas y yacimientos, longitud de ARPSIs y superficies inundables, número de habitantes por municipio y por zona inundable y número de amarres. Ninguna de las variables mencionadas ha sido formada por una asignación secuencial ni de manera aleatoria, los datos no tienen valores acotados y el tamaño de la muestra es superior a 100 datos. Estas características hacen que las variables sigan la distribución del primer dígito definida por Benford. Una de las condiciones más restrictivas es el rango de valores, pues todas las variables cuyos valores están acotados (altura de ola, periodo de pico, marea meteorológica, marea astronómica y densidad de población) no cumplen la Ley de Benford. Con ello, se plantea la posibilidad de emplear esta herramienta para la detección de errores o manipulación de los conjuntos de datos cuyas características son conformes con la Ley de Benford.

ABSTRACT

A common pattern in the first-digit frequency distribution of different sets of real numbers opens the possibility of demonstrating the consistency of a database through that distribution. This pattern, defined by Benford's Law, considers digits 1, 2 and 3 more likely, with a decrease in the frequency of occurrence to 9. In this study, a database is submitted with both physical and socioeconomic variables of the coastal area and the conformity of those variables that meet the conditions for the application of the Benford Act is verified. This database comes from the National Strategic Plan for the Protection of the Spanish Coast considering the Effects of Climate Change and has information on the entire Spanish coast. It extracts variables of interest for coastal management related to beaches, watersheds, maritime climate, underwater deposits, potentially floodable areas, population and moorings in marinas. These variables undergo goodness-of-fit tests that determine compliance with the Benford Act. The results show a high conformity for the length of beaches, erosion stretches on beaches, area of basins and deposits, length of ARPSIs and flood areas, number of inhabitants per municipality and flood zone and number of moorings. None of the variables mentioned have been formed by a sequential or random assignment, the data have no bounded values and the sample size is greater than 100 data. These characteristics cause the variables to follow the first digit distribution defined by Benford. One of the most restrictive conditions is the range of values, since all variables whose values are bounded (wave height, peak period, meteorological tide, astronomical tide and population density) do not comply with the Benford Law. This raises the possibility of using this tool for the detection of errors or manipulation of data sets whose characteristics comply with the Benford Act.

1. INTRODUCCIÓN

Los estudios en la costa, así como en cualquier otro ámbito, se fundamentan en las bases de datos. El análisis estadístico que permite determinar patrones y tendencias, requiere una cierta calidad de los conjuntos de datos para que los resultados sean representativos. Cuando se habla de calidad de un conjunto de datos, se hace referencia a la exactitud, completitud, integridad, actualización, coherencia, relevancia y confiabilidad (Yunta, 1998). Sin embargo, generar bases de datos con estas características es una tarea complicada. Por ello, es importante testar estas bases de datos y conocer sus limitaciones para inferir en aquellas variables que puedan condicionar negativamente los resultados de un análisis estadístico.

Para validar la calidad de cualquier conjunto de datos es necesario encontrar una característica común o un patrón que sea independiente del contenido de la base de datos. Así pues, se introduce, como posible control de calidad, una ley matemática que define el comportamiento del primer dígito de cada dato, la Ley de Benford (1938). Esta ley afirma que la distribución de los primeros dígitos de un conjunto de datos es predecible. Así pues, la frecuencia de aparición del primer dígito disminuye del 1 al 9. Benford afirma que el 30% de los números de un conjunto de datos posee como primer dígito significativo el 1. Esta probabilidad desciende a un 17.6% para el dígito 2 y se produce un decrecimiento gradual hasta el dígito 9 con un 4.5% de probabilidad.

Entre otras aplicaciones, estudios como Nigrini y Miller en 2007 y Geyer y Martí en 2012, tratan de validar la calidad de las bases de datos de ciertos campos de las ciencias de la tierra mediante la Ley de Benford. Nigrini y Miller (2007) aplican a Ley de Benford a datos de hidrología, concretamente datos de lagos y humedales, los cuales se demostró que no siguen esa distribución del primer dígito tal y como la define Benford. Por el contrario, los resultados de Geyer y Martí (2012), en su análisis de datos referidos a la vulcanología, si se ajustan a la Ley de Benford y permiten detectar un exceso de redondeo de datos, errores o anomalías.

Siguiendo la línea de estos estudios, el objetivo del presente trabajo es, en primer lugar, comprobar si las distintas variables del ámbito costero se ajustan a la Ley de Benford. En segundo lugar, definir los límites de aplicación de la ley y, por último, en función de estos límites, evaluar la utilidad potencial de la Ley de Benford como control de calidad de las bases de datos en el ámbito costero.

A diferencia de los estudios mencionados, este trabajo analiza conjuntos de datos de variables tanto físicas, como socioeconómicas. Cada conjunto posee características distintas, lo que permite explorar en mayor medida los límites de esta ley matemática.

Todas las variables empleadas se encuentran agregadas en el Plan Estratégico Nacional para la Protección de la Costa Española considerando los Efectos del Cambio Climático. Así pues, se dispone de una base de datos con información de todo el litoral español.

2. LEY DE BENFORD

En el presente apartado se define de manera breve el origen de la Ley de Benford y como adquirió repercusión con el paso del tiempo hasta las aplicaciones de los últimos años. Seguidamente, se muestran las bases teóricas de dicha ley matemática y algunas de las condiciones para el cumplimiento de la misma, que han sido recopiladas mediante una revisión bibliográfica.

En el siglo XIX, Simon Newcomb observó que, en las tablas de logaritmos, las primeras páginas se encontraban más desgastadas por el uso. Pues bien, las tablas de logaritmos están ordenadas de tal manera que, las primeras páginas corresponden con los números cuya primera cifra es el 1 y las últimas con aquellos cuyo primer dígito es el 9. Con ello, Newcomb dedujo que los números con mayor ocurrencia en cálculos cotidianos tenían como primer dígito el número 1 y dicha probabilidad decrecía hasta el 9. Esta hipótesis fue publicada en 1881 en la revista *American Journal of Mathematics* bajo el título “*Note on the Frequency of Use of the Different Digits in Natural Number*”. Sin embargo, el artículo no mostró gran aceptación por parte de la comunidad matemática debido a la falta de comprobación empírica de la ley.

No fue hasta el año 1938, cuando Frank Benford detecta (sin conocimiento previo acerca del tema) la misma anomalía en las tablas de logaritmos. A diferencia de Newcomb, Benford comprobó empíricamente lo que denominó Ley de los número anómalos mediante la recolección de hasta 20.229 números de distintas fuentes. A través de periódicos, longitudes de ríos, tablas de constantes matemáticas y físicas, estadísticas deportivas, etc., Benford calculó la probabilidad de aparición del primer dígito significativo de cada conjunto de datos. Los resultados de su estudio fueron publicados en *Proceedings of the American Philosophical Society*, con una mayor repercusión. Fueron varios los autores que tras esta publicación trataron de elaborar artículos ampliando el contenido de la materia. Alguno de ellos, como Ralph A Raimi en 1989, descategorizaba la ley por falta de fines prácticos. Esta carencia se vincula con la falta de herramientas informáticas en la época, que dificultaba la comprobación empírica de la ley, ya que, como se explicará más adelante, el cumplimiento de la Ley de Benford requiere procesar una cantidad de datos elevada.

Es Mark Nigrini, en el año 1991, quien encuentra una aplicación en la detección de anomalías en las declaraciones de impuestos. De tal manera que, si la Ley de Benford se cumple en bases de datos naturales, se podría identificar si existe manipulación en ellas al diferir de las frecuencias de aparición esperadas. Al atribuirle un fin práctico, la Ley de Benford comienza a tener más repercusión. Tras

ello, surgen nuevas aplicaciones, pero no es hasta 1995 cuando Ted Hill aporta una prueba matemática rigurosa de la ley (Teorema del Límite Central).

La suma de evidencias arrojadas por dichos autores, entre otros, abren un amplio abanico de aplicación de la Ley de Benford. La mayor parte de los estudios pertenecen al ámbito financiero (Durtschi et al. 2004, Tödter et al. 2009, Deckert et al. 2011). No obstante, se ha experimentado también en el campo de la física, la astronomía, la geofísica, la química, la ingeniería y las matemáticas (Dumas et al. 2000, Nigrini et al. 2007, Sambridge et al. 2010). Así pues, la revisión bibliográfica evidencia una amplia variedad de aplicaciones como COVID-19, fraudes electorales, redes neuronales, etc. Algunos autores ya mencionados, como Nigrini y Miller en 2007 y Geyer y Martí en 2012, además de aplicar la ley a ciertos campos de las ciencias de la tierra, tratan de comprobar si la Ley de Benford podría tener una utilidad en estos campos como herramienta para la detección de inconsistencia de las bases de datos.

Así pues, la Ley de Benford, también denominada ley del primer dígito o ley de los números anómalos, establece que la frecuencia de aparición del primer dígito significativo sigue un patrón predecible. Esta ley sustenta que la distribución de dígitos en números naturales no es aleatoria, sino que sigue la probabilidad de ocurrencia que se muestra en la Tabla 1 dando como resultado una curva logarítmica (Figura 1). Cabe mencionar que el cero no se considera primer dígito significativo.

Tabla 1. Proporción del primer dígito según la Ley de Benford

Primer dígito	Probabilidad de ocurrencia
0	-
1	0.30103
2	0.17609
3	0.12494
4	0.09691
5	0.07918
6	0.06695
7	0.05799
8	0.05115
9	0.04576

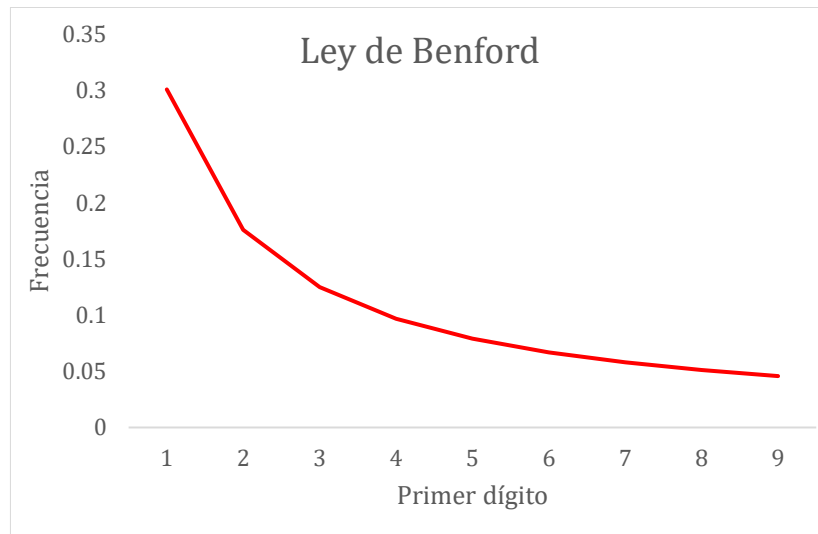


Figura 1. Distribución de frecuencias del primer dígito según la Ley de Benford.

Estos valores de probabilidad de cada dígito vienen dados por la siguiente expresión:

$$Prob(d_1) = \sum \log_{10}\left(1 + \frac{1}{d_1}\right) \quad (\text{Ec.1})$$

Donde d_1 es el primer dígito significativo distinto de cero.

Se realiza una revisión bibliográfica acerca de los criterios que favorecen que una variable se ajuste a la Ley de Benford. A continuación, se definen las condiciones de mejor ajuste según estudios previos:

- La ocurrencia natural de las colecciones de números. Es decir, que los datos no hayan sido asignados secuencialmente o influenciados por el pensamiento humano, ni tampoco que sean producto de un proceso aleatorio (Benford, 1938).
- Cada conjunto de datos debe estar compuesto por magnitudes medibles de un mismo fenómeno.
- Valores distribuidos en varios órdenes de magnitud. Cuantos más órdenes de magnitud cubran los datos de manera uniforme, mayor será el ajuste (Nigrini, 1997).
- El tamaño de la muestra, a pesar de ser un factor importante para el ajuste, no está definido en el estado del arte. Existen discrepancias entre autores que establecen el límite inferior en 100 datos como Fewster en 2009, frente a autores que aseguran un buen ajuste para la Ley de Benford en conjuntos de datos superiores a 1000 (Nigrini, 1997).

- Bases de datos no manipuladas en cuanto al redondeo en la expresión decimal (Nigrini, 2012).
- Distribución del primer dígito invariante ante los cambios de escala e independiente de las unidades de medida (Pinkham, 1961)
- Misma tecnología de medición para la totalidad del conjunto de datos (Nigrini, 2007).

Sin embargo, estos criterios de mejor ajuste no son concluyentes a la hora de determinar si una variable aplica o no a la Ley de Benford, el estado del arte evidencia discrepancias en dichas restricciones. Además, no se trata de condiciones que establezcan un límite terminante, sino que la conformidad con la ley aumenta o disminuye gradualmente.

3. BASE DE DATOS

En el presente apartado se define y caracteriza la base de datos, analizando cada variable de manera individual.

3.1. Fuente de datos

Todos los datos empleados en este estudio proceden del Plan Estratégico Nacional para la Protección de la Costa Española considerando los Efectos del Cambio Climático. Se trata de un proyecto requerido por el Ministerio de Transición Ecológica y el Reto demográfico (MITECO) y realizado por The Coastal & Marine Union (EUCC), el Instituto de Hidráulica Ambiental (IHCantabria) y la Universidad de Las Palmas de Gran Canaria (ULPGC). La base de datos usada en dicho proyecto ha sido recopilada por el IHCantabria, y en ella se incluyen variables físicas, socioeconómicas y medioambientales de la costa española. De la totalidad de datos que contiene esta base de datos, se han seleccionado aquellas variables cuantitativas, con especial interés en los estudios de gestión de la costa.

Los datos de partida provienen, por un lado, de una revisión bibliográfica a escala nacional y, por otro lado, de Consultas Técnicas al personal de la Dirección General de la Costa y el Mar (DGCM) a nivel provincial.

El propio MITECO ofrece una extensa base de datos físicos del litoral. En primer lugar, tiene a disposición la Guía de Playas que proporciona información de la longitud de las playas que se diferencian a lo largo de la costa española. En segundo lugar, ofrece información acerca de la superficie de las cuencas y subcuencas hidrográficas a escala 1:25.000. Por último, dispone de un Visor de Información Geográfica Marina en el ámbito de los Planes de Ordenación del Espacio Marítimo (OEM) por la Dirección General de la Costa y el Mar (DGCM), del cual se puede obtener información como la valoración de la extensión de los yacimientos submarinos de arena. Esta fuente de datos no posee la misma fecha de actualización

para el conjunto de datos ya que aún está en proceso de elaboración. Además, presenta niveles de detalle diversos.

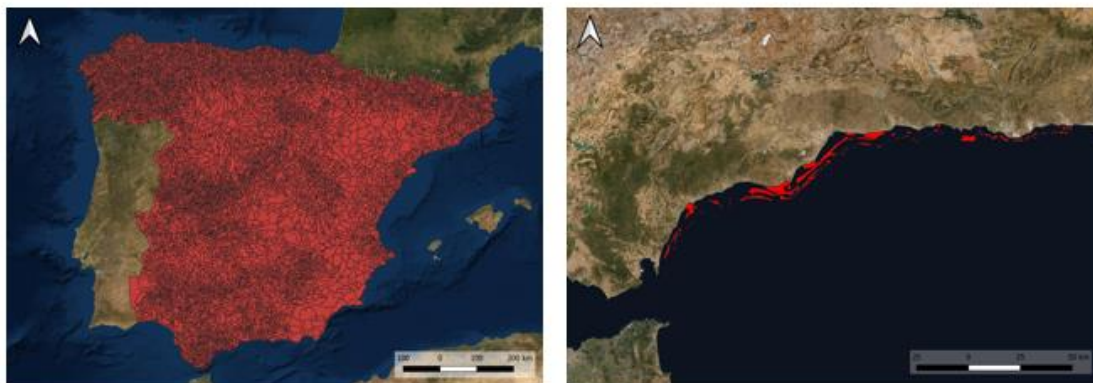


Figura 2. A: Cuencas y subcuencas hidrográficas de la península ibérica. B: Representación de la superficie de yacimientos submarinos en la costa de Andalucía.

En cuanto a la información de clima marítimo, ésta proviene del IHData, base de datos desarrollada por el IHCantabria que proporciona información histórica de oleaje y nivel del mar desde 1948 hasta la actualidad a partir de series de reanálisis (DOW, GOS y GOT) con una resolución horaria. Se cuenta con datos de marea meteorológica (GOS), marea astronómica (GOT) y olas en la costa (DOW).



Figura 3. Localización de los puntos de reanálisis con información de clima marítimo en la costa gallega.

Otra variable física que contempla el informe, son las Áreas con Riesgo Potencial Significativo de Inundación (ARPSIS, 2º Ciclo) que representan las áreas amenazadas por inundación de origen marino y marino-fluvial, en función de vulnerabilidad de la zona, bajo criterios medioambientales, presencia de población y actividades económicas. Dicha información proviene del Sistema Nacional de Cartografía de Zonas Inundables (SNCZI).



Figura 4. Representación de la extensión de las ARPSIs en la Isla de Gran Canaria

A través del censo municipal del Instituto Nacional de Estadística (INE) se obtienen datos de población en cuanto a número y densidad. Para poder agregar dicha información se obtienen límites municipales y provinciales del Centro de Descargas del Centro Nacional de Información Geográfica. La posible falta de precisión en las mediciones de estas áreas y la posterior digitalización hacen que estos archivos puedan tener hasta 40 m de incertidumbre.

Se cuenta, además, con información del número de amarres ligado a cada puerto deportivo existente en la costa española. Estos datos provienen de la web Chárter fondear.

Como se menciona anteriormente, otra serie de datos proceden de consultas técnicas realizadas, principalmente, al personal técnico de las demarcaciones de costas y servicios provinciales. Los cuestionarios realizados arrojan información acerca de la erosión de las playas, entre otras temáticas. Con ayuda de identificadores y datos geográficos, el personal técnico estima el porcentaje de playa que experimenta erosión, tanto crónica como eventual.

3.2. Clasificación de las variables

El Plan Estratégico Nacional para la Protección de la Costa Española abarca información socioeconómica y física. No obstante, esta clasificación no resulta del todo exacta para algunas variables. Los yacimientos submarinos son depósitos de arena definidos y localizados por el ser humano con fines de explotación, por lo cual podrían considerarse a su vez variables socioeconómicas. De igual manera, las ARPSIs, como se mencionaba anteriormente, se definen en función de la vulnerabilidad y exposición de la zona atendiendo a riesgos económicos y de población. Cabe mencionar que, las variables vinculadas a unidades de ARPSIs (superficies de inundación y habitantes por zona inundada, también están sujetas a este criterio.

A continuación, se presenta una tabla con las variables seleccionadas para el estudio atendiendo a su tipología.

Tabla 2. Clasificación del tipo de variables empleadas en el estudio.

Variables físicas
Longitud de playas
Erosión crónica
Eventos de erosión
Área de cuencas
Altura de ola
Periodo de pico
Marea astronómica
Marea meteorológica
Variables físicas/socioeconómicas
Área de yacimientos
Longitud de ARPSIs
Superficie inundación
Variables socioeconómicas
Habitantes/zonas inundables
Habitantes/municipio
Densidad de población
Número de amarres

3.3. Descripción de los datos

A continuación, se definen las características de cada conjunto de datos que puedan condicionar el cumplimiento de la Ley de Benford según algunas de las condiciones enumeradas en el Apartado 2.

- Longitud de playas

La longitud de las playas es una variable natural expresada en metros. Sin embargo, tiene un matiz humano en la división de la unidad de análisis. Se asocia un valor de longitud a una playa individual, pero en ocasiones, se dividen playas que físicamente es una unidad. El tamaño de la muestra es de 5335 playas. Los valores de longitud están expresados en metros y su expresión decimal ha sido eliminada, es decir, hay cierto error en la medida debido al redondeo. La diferencia entre la longitud mínima y máxima es de tres órdenes de magnitud.

- Erosión Crónica y Eventual de las playas

A partir de la longitud de playa y los porcentajes de erosión que se obtuvieron en las encuestas realizadas se calcula la longitud de playa afectada por la erosión crónica y los eventos de erosión. Estas variables están condicionadas por el criterio

del personal técnico, común para cada provincia. El volumen de datos difiere del número de playas al eliminar los resultados nulos, aquellas playas que no están sujetas a erosión. La muestra de erosión crónica cuenta con 1061 datos frente a 1080 playas afectadas por los eventos de erosión. Las zonas afectadas por la erosión crónica se distribuyen por tres órdenes de magnitud, mientras que los eventos erosivos resultan un poco más acotados, dos órdenes de magnitud. Los valores no están redondeados, pero están calculados a partir de la longitud de playa que, como se mencionaba, sí son valores redondeados.

- Área de las cuencas hidrográficas

Este conjunto de datos es una variable natural que, a diferencia de las playas, se trata de una unidad indivisible, ya que se define como la superficie de terreno cuya escorrentía superficial fluye en su totalidad por una única desembocadura, estuario o delta. El tamaño de la muestra es de 13424 cuencas, con valores de superficie expresados en m^2 y redondeados. La diferencia entre el área más pequeña y el más grande abarca seis órdenes de magnitud.

- Clima marítimo

Los parámetros que definen el clima marítimo (altura de ola (H_s), periodo de pico (T_p), marea meteorológica (MM) y astronómica (MA)) provienen de estadísticas climatológicas con una precisión similar a los datos instrumentales. Por ello, se pueden considerar variables naturales. Son conjuntos de datos de amplia extensión, prolongados en el tiempo y no afectados por el redondeo. El tamaño de la muestra de las 4 variables es de 582894. Sin embargo, su potencial de cumplimiento de la Ley de Benford se reduce cuando se habla del rango de valores, es decir, la diferencia entre el máximo y mínimo. Así pues, todas ellas son variables acotadas, que se distribuyen en un rango de valores pequeño.

- Área de los yacimientos submarinos

Este conjunto de datos cuenta con 557 datos expresados en m^2 , los cuales no están redondeados. El rango de valores de estas superficies se extiende por 10 órdenes de magnitud.

Como se define durante la clasificación de las variables en el Apartado 3.2., los yacimientos submarinos, así como las ARPSIs, que se muestran a continuación, y aquellas variables sujetas a esta unidad de análisis, son parámetros físicos delimitados por el ser humano. A pesar de que los valores de estas variables no tengan una ocurrencia natural, tampoco han sido asignados de manera secuencial ni de manera aleatoria, sino que, cada conjunto de datos atiende a unos criterios específicos, mencionados anteriormente en el Apartado 3.1.

- ARPSIs

La base de datos cuenta con 2913 unidades de ARPSIs, cuya longitud, en metros, no está redondeada y abarca cuatro órdenes de magnitud. A partir de estos datos, se obtienen láminas de inundación para periodos de retorno de 100 y 500 años.

Estos conjuntos de datos tienen un menor tamaño de muestra ya que no contemplan la inundación fluvial, solo la marina. Así pues, se reduce el número de ARPSIs a 398 para ambos periodos de retorno y se expresa la extensión de la superficie inundable en metros, con valores redondeados de cuatro órdenes de magnitud diferente.

Otra variable que surge a partir de las ARPSIS es el número de habitantes por zona inundable. Mediante las láminas de superficie inundable se calcula la cantidad de población en riesgo, expresada en número de habitantes por área. La variación de estos datos difiere hasta 4 órdenes de magnitud para ambos periodos de retorno.

- Número de habitantes por municipio costero y densidad de población por m^2

Si se toma los datos de población en número de habitantes por municipio costero, se cuenta con 497 datos que varían hasta cuatro órdenes de magnitud. Sin embargo, si se expresa en densidad de población, es decir, en habitantes por m^2 , los valores están acotados a un único orden de magnitud.

- Número de amarres por puerto deportivo

Por último, la base de datos cuenta con 395 puertos deportivos, a partir de los cuales se contabilizó el número de amarres. Nuevamente, a pesar de no ser una variable natural, estos valores no han sido asignados de manera secuencial ni aleatoria. El rango de valores se distribuye por tres órdenes de magnitud.

4. ANÁLISIS DE LOS DATOS

4.1. Metodología de análisis

Inicialmente, se realiza el mismo procedimiento para todos los conjuntos de datos, se extrae el primer dígito significativo distinto de cero y se calcula la frecuencia de aparición de cada dígito del 1 al 9. En este primer ejercicio, se trabajan los conjuntos de datos desagregados, es decir, la unidad de análisis más pequeña que permita aumentar el volumen de datos de cada variable.

Una vez se obtiene la probabilidad de ocurrencia de cada dígito, se someten los resultados a una prueba de bondad de ajuste. Mediante una revisión bibliográfica se selecciona el test de la Desviación Media Absoluta (MAD) que proporciona mayor rigurosidad al análisis (Nigrini, 2012). Además, a diferencia de otros métodos, MAD no se ve limitado por la cantidad de datos. Este test estadístico se define mediante la siguiente expresión:

$$MAD = \frac{1}{9} \sum [P_{obs}(d) - P_t(d)] \quad (Ec.2)$$

Donde $P_t(d)$ es la proporción esperada según Benford y P_{obs} es la proporción observada.

Morales en 2020, establece unos valores críticos de MAD que determinan la conformidad o no conformidad con la Ley de Benford. Se muestran los rangos a continuación, en la Tabla 3:

Tabla 3. Rangos de MAD para definir el nivel de conformidad

Rango	Nivel de Conformidad
< 0.016	Alto
0.016 – 0.03	Aceptable
0.03 – 0.06	Marginalmente aceptable
>0.06	Bajo

Los niveles de conformidad resultantes para cada variable permiten realizar una valoración de las limitaciones que se encuentran a la hora de aplicar la Ley de Benford. Para ello, se modifican algunos conjuntos de datos variando aquellas condiciones que limitan el cumplimiento de la ley, como el tamaño de la muestra, las unidades de medida, el rango de valores, etc.

4.2. Resultados del análisis

Variables físicas

La longitud de las playas es una variable con un alto potencial de ajuste a la Ley de Benford como se describe en el Apartado 3.3. Así pues, en la Figura 5 se puede apreciar la similitud entre la probabilidad de ocurrencia del primer dígito para la variable en cuestión y la probabilidad definida por Benford. Como era de esperar, la prueba de bondad de ajuste muestra un nivel de conformidad alto con una desviación media absoluta de 0.007. Esta pequeña desviación de la curva puede estar sujeta a la incertidumbre generada por el redondeo de los datos o la división de playas que físicamente son la misma.

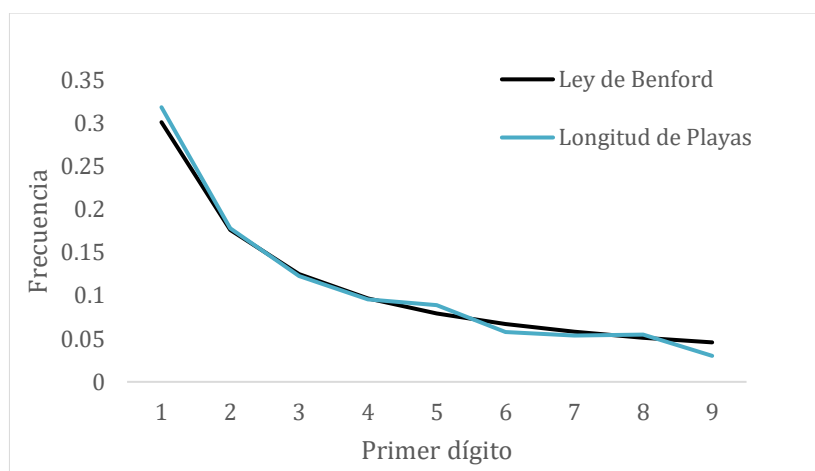


Figura 5. Proporciones del primer dígito para la longitud de playa frente a la Ley de Benford.

Las variables que representan el tramo de dichas playas afectado por la erosión también adquieren un nivel de conformidad alto, con una MAD de 0.0093 para la erosión crónica y 0.007 para los eventos de erosión. Este resultado llama la atención por la fuente a la que están sujetos. Como se menciona anteriormente, estos valores de erosión son el resultado de porcentajes estimados por el personal técnico, es decir, la precisión de los datos no es del todo rigurosa. Las probabilidades del primer dígito son semejantes en los dos conjuntos de datos, llegando a coincidir en los dígitos 4, 6 y 8 (Figura 6).

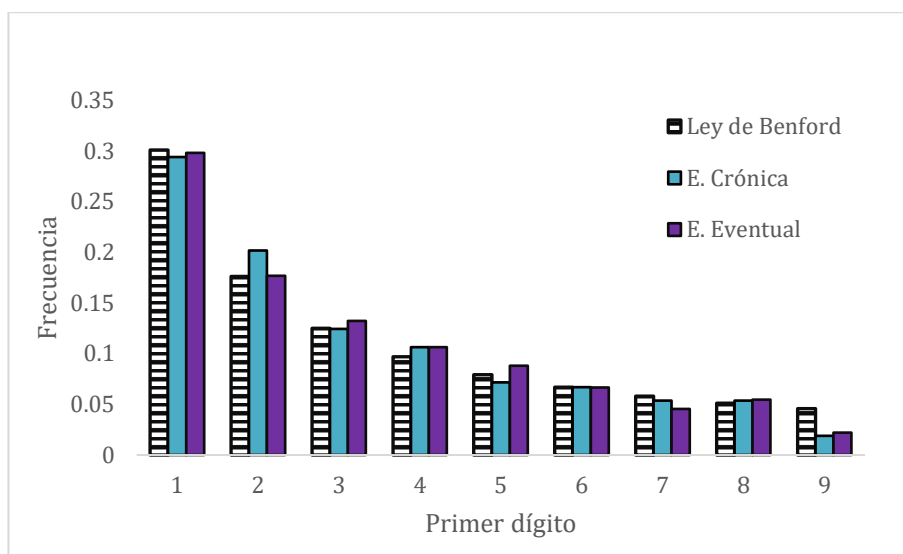


Figura 6. Proporciones del primer dígito para los valores de erosión de playa frente a la Ley de Benford.

Al igual que la longitud de las playas, otra de las variables con un alto potencial de ajuste es el área de las cuencas hidrográficas. Así pues, la probabilidad del primer dígito decrece del 1 al 9. Sin embargo, y a pesar de que el tamaño de la muestra y el rango de valores es mayor se puede apreciar en la desviación de la curva en la Figura 7 que la MAD es superior (0.016), situándose en el límite entre un nivel de conformidad alto y aceptable. Nuevamente se trata de datos redondeados.

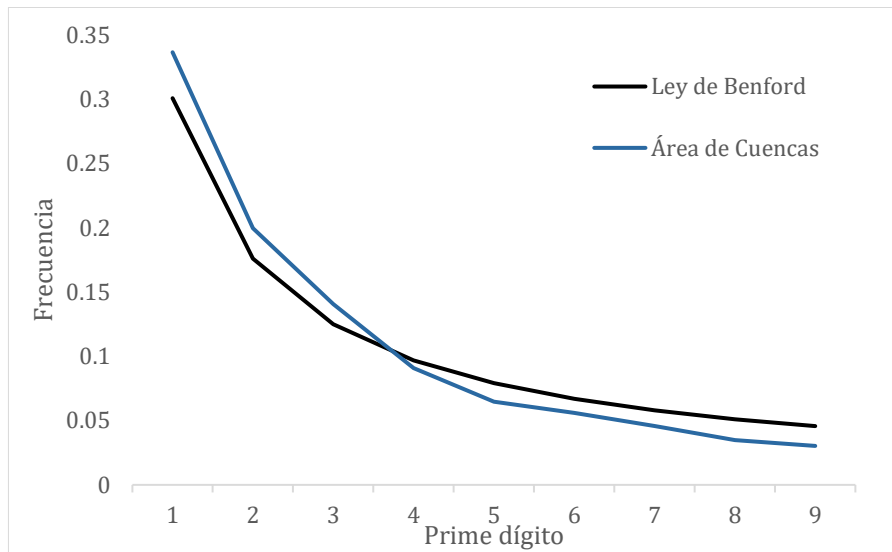


Figura 7. Proporciones del primer dígito para el área de las cuencas frente a la Ley de Benford.

A diferencia de las anteriores variables, no se espera un buen ajuste por parte de los datos de clima marítimo. El análisis de Hs muestra una mayor probabilidad de los dígitos 1 y 2 (Figura 8). Este resultado es coherente si se observa la serie de datos (Figura 9) donde se aprecia una mayor ocurrencia de alturas de ola entre 1 y 2 metros. El punto de reanálisis que se ha empleado para el análisis corresponde con una zona donde las olas superiores a 4 m son menos recurrentes, por ello la probabilidad del que el primer dígito este entre el 4 y 9 es baja. Si se hubiese tomado un punto de reanálisis de un lugar más energético, donde las olas fuesen de mayor tamaño, esta distribución cambiaría.

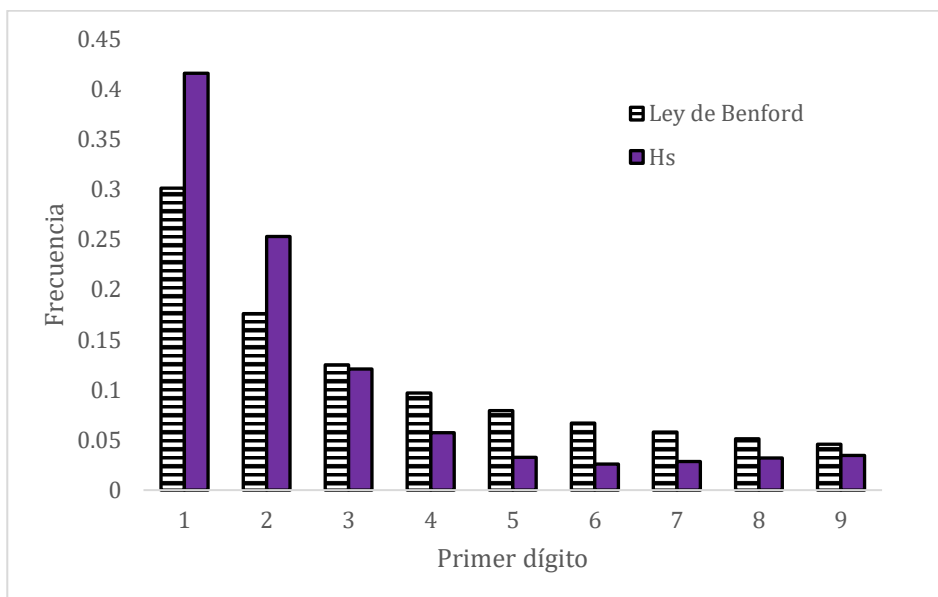


Figura 8. Proporciones del primer dígito para la altura de ola frente a la Ley de Benford

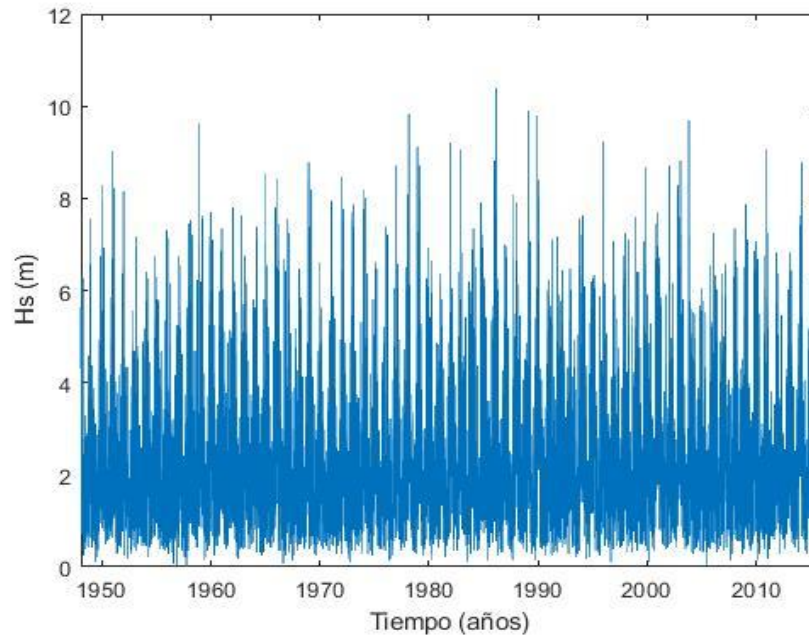


Figura 9. Serie temporal de altura de ola.

De igual manera, el periodo de pico se encuentra en un rango de valores (Figura 10) que se ve reflejado en la distribución del primer dígito (Figura 11). Los periodos más frecuentes en este punto de reanálisis están comprendidos entre los 6 y 14 s, por ello, que el primer dígito este entre el 2 y el 5 es prácticamente improbable.

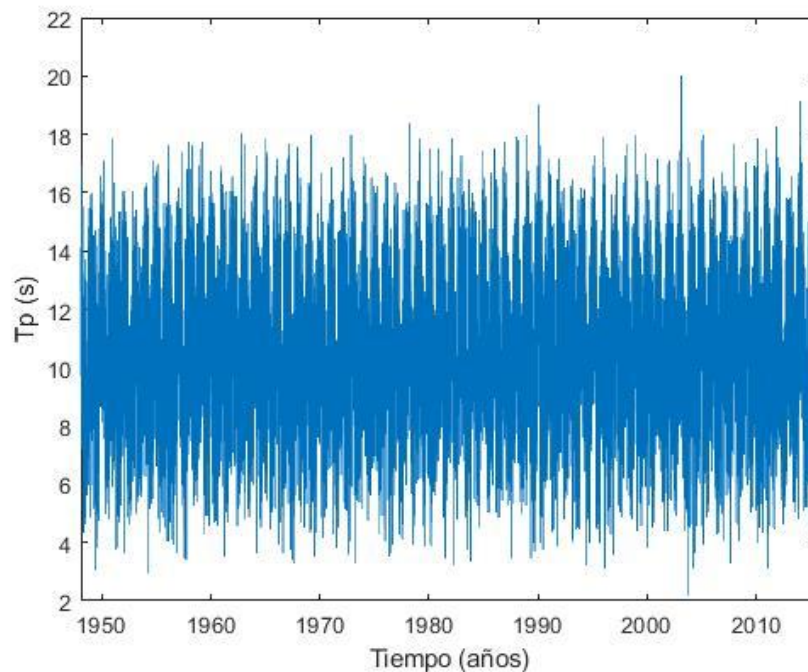


Figura 10. Serie temporal de periodo de pico.

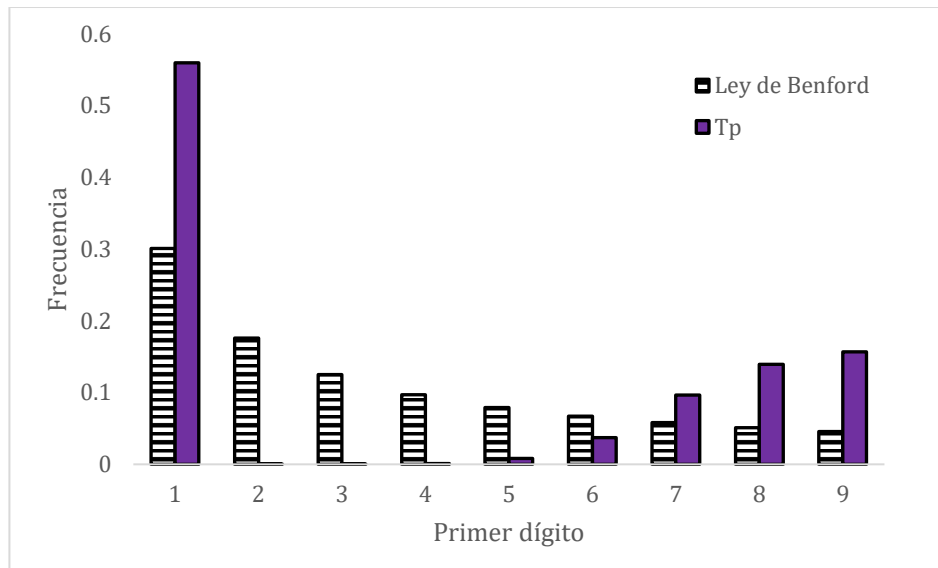


Figura 11. Proporciones del primer dígito para el periodo de pico frente a la Ley de Benford.

La marea meteorológica del punto seleccionado varía entre el 0 y 0.2 m (Figura 12). Como el cero no se considera primer dígito significativo, los primeros dígitos con mayor frecuencia son el 1 y el 2 y 3 en menor medida como se puede apreciar en la Figura 13.

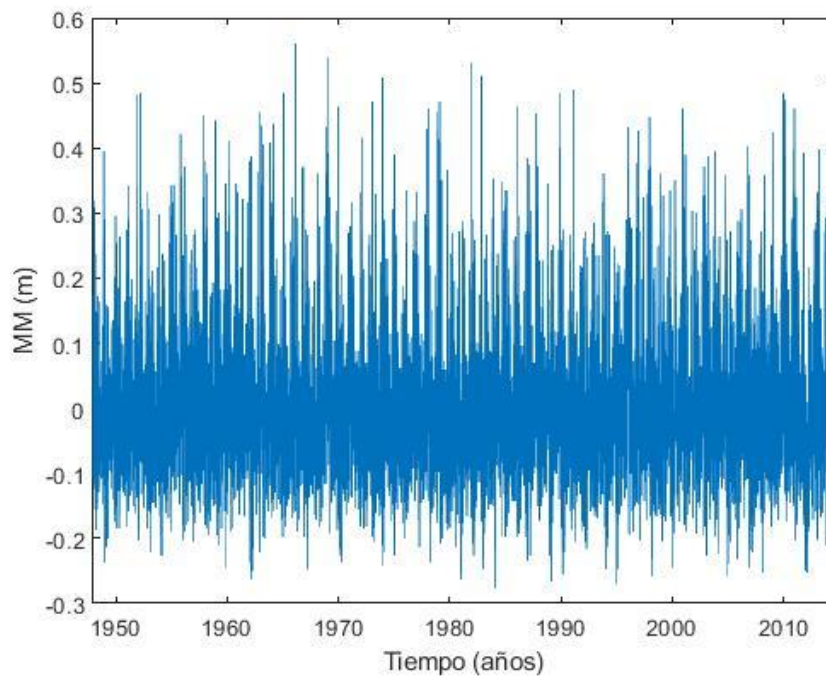


Figura 12. Serie temporal de marea meteorológica.

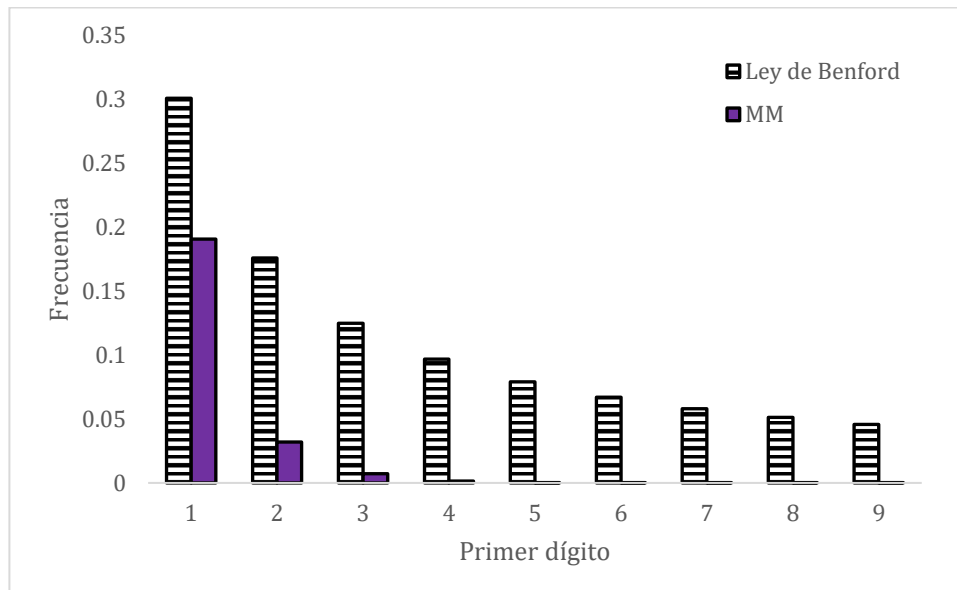


Figura 13. Proporciones del primer dígito para la marea meteorológica frente a la Ley de Benford

En cambio, la marea astronómica se distribuye por todos los dígitos. El 1 es el más frecuente y del 2 al 5 la probabilidad es más o menos uniforme (Figura 14). A diferencia de la marea meteorológica, la marea astronómica es una variable determinista cuyo comportamiento es periódico y predecible debido a las interacciones gravitacionales entre la Tierra, la Luna y el Sol. El rango de marea es mayor que la meteorológica, casi 2 m (Figura 15), lo que permite que el primer dígito significativo varíe entre el 1 y el 9. No obstante, sigue siendo una variable muy acotada bajo los criterios de la Ley de Benford.

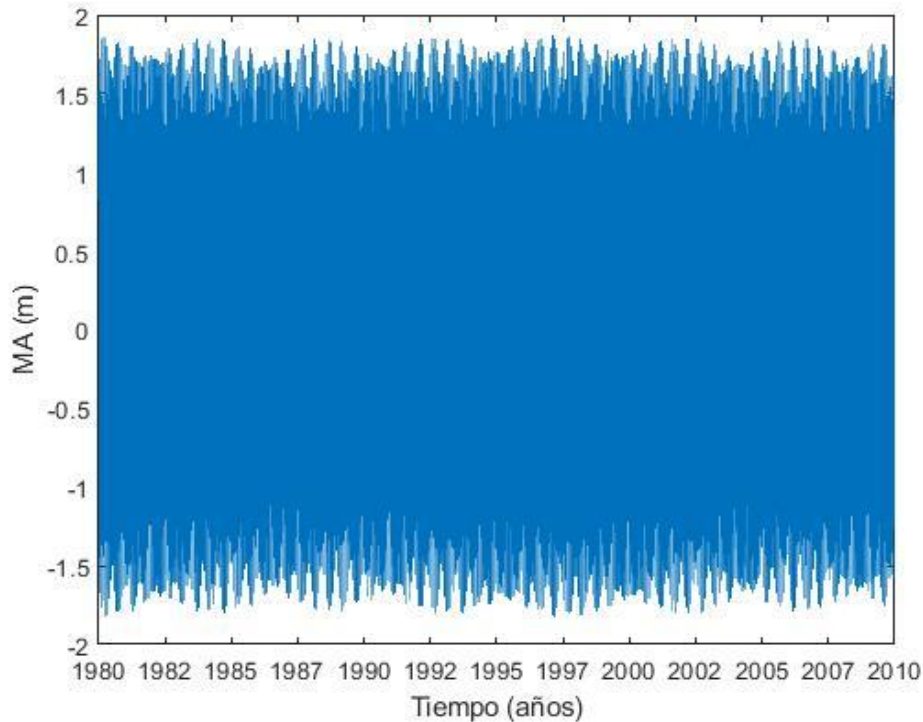


Figura 14. Serie temporal de marea astronómica.

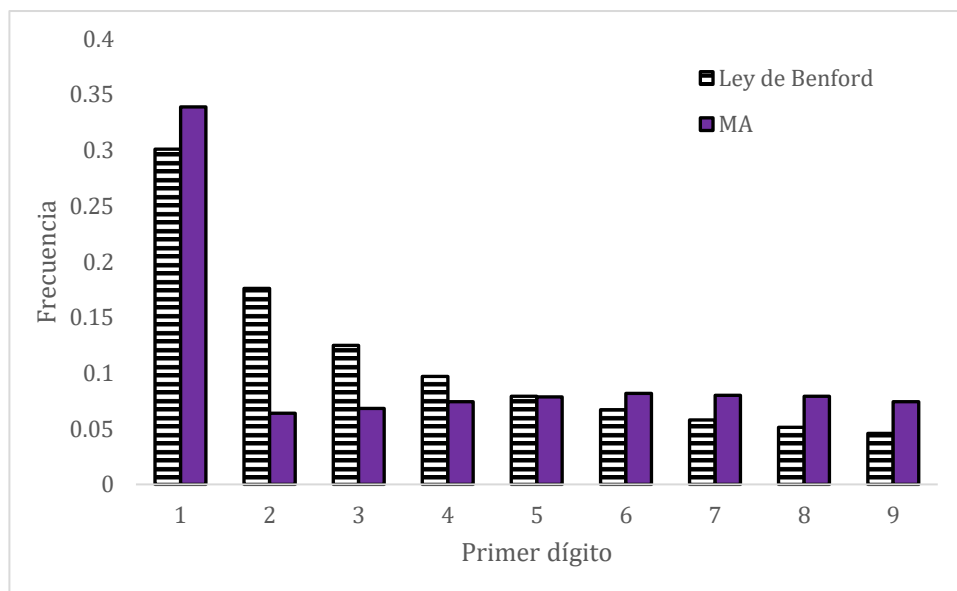


Figura 15. Proporciones del primer dígito para la marea astronómica frente a la Ley de Benford.

Variables físicas/socioeconómicas

En teoría el potencial de cumplimiento del conjunto de datos de yacimientos es inferior al de playas y cuencas. Como se menciona en el anterior apartado, el área de los yacimientos tiene una fuerte componente humana. Además, la muestra no

supera los 600 datos. Cabe recordar también que se trata de una fuente con distintas fechas de actualización e incompleta. A pesar de estas características desfavorables para el cumplimiento, el nivel de conformidad para esta variable es alto (Figura 16), con una MAD de 0.011. De la igual manera, la longitud de las ARPSIs, también dependiente del criterio humano, obtiene un nivel de conformidad alto (0.012). Con una pequeña desviación, la variable se aproxima bastante a la curva de Benford (Figura 17).

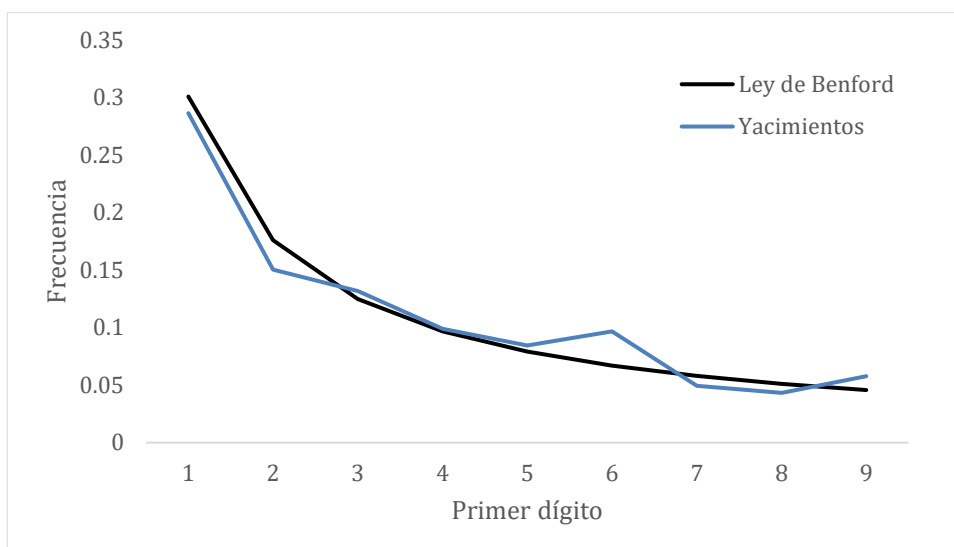


Figura 16. Proporciones del primer dígito para el área de los yacimientos frente a la Ley de Benford.

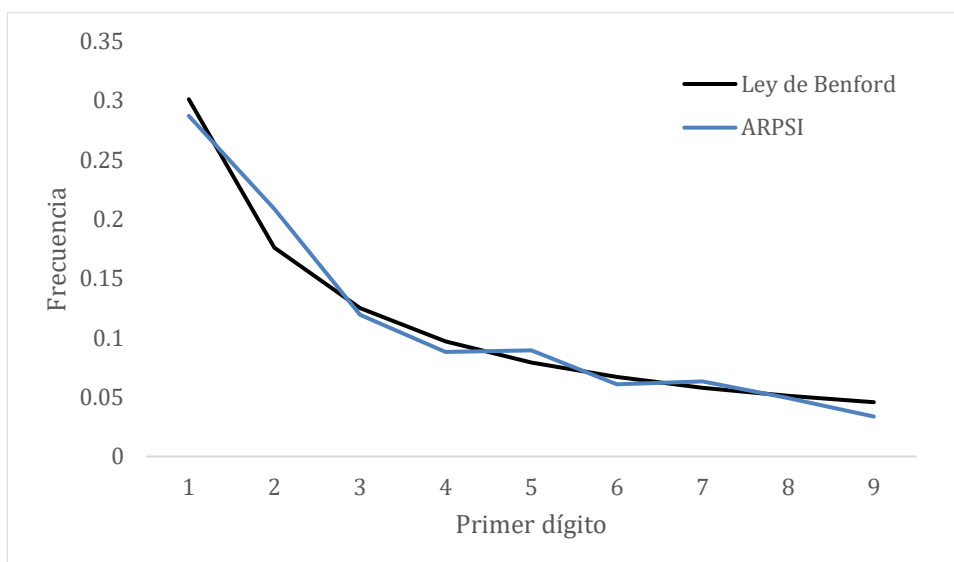


Figura 17. Proporciones del primer dígito para la longitud de las ARPSIs frente a la Ley de Benford.

Las láminas de inundación, obtenidas a través de las ARPSIs para periodos de retorno de 100 y 500, también se adecúan, en cierta medida, a la Ley de Benford

(Figura 18). Se puede diferenciar un nivel de conformidad inferior para el periodo de retorno de 100 años con una MAD de 0.018 (aceptable), frente a 0.015 para un periodo de retorno de 500 años (alto). No obstante, esta diferencia de la desviación media absoluta no es lo suficientemente grande como para poder afirmar una dependencia de la conformidad con el periodo de retorno. A diferencia del dígito 6, los demás dígitos siguen el mismo patrón de desviación para ambos periodos de retorno, incluso se igualan las frecuencias para los dígitos 8 y 9.

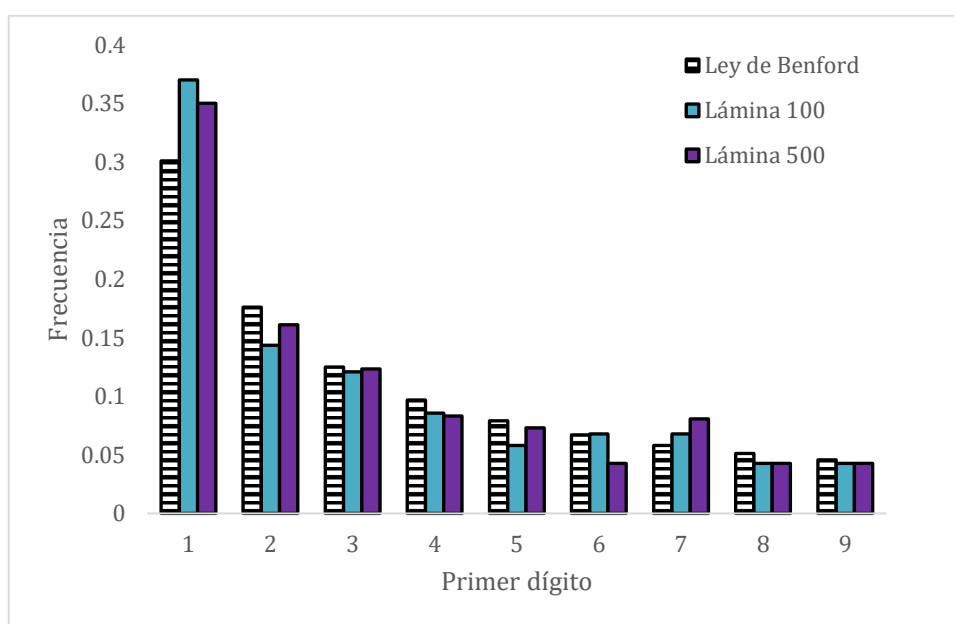


Figura 18. Proporciones del primer dígito para las superficies inundables frente a la Ley de Benford.

Si se analiza el número de habitantes en esas superficies nuevamente se obtiene un mejor ajuste para el periodo de retorno de 100 años. En este caso, ambos conjuntos de datos presentan un nivel de conformidad alto con una MAD de 0.009 para los 100 años y 0.008 para un periodo de retorno de 500 años, nuevamente la diferencia es ínfima como para justificar un vínculo entre la conformidad y el periodo de retorno.

Variables socioeconómicas

Por último, se analizan las variables socioeconómicas, el número de habitantes por municipio costero y el número de amarres por puerto deportivo. Esta última, no sigue un patrón decreciente de la probabilidad del 1 al 9, sino que la probabilidad de ocurrencia del dígito 6 supera al 4 y al 5 y el 8 supera al 7 (Figura 19). No obstante, presenta un nivel de conformidad aceptable con una MAD de 0.022.

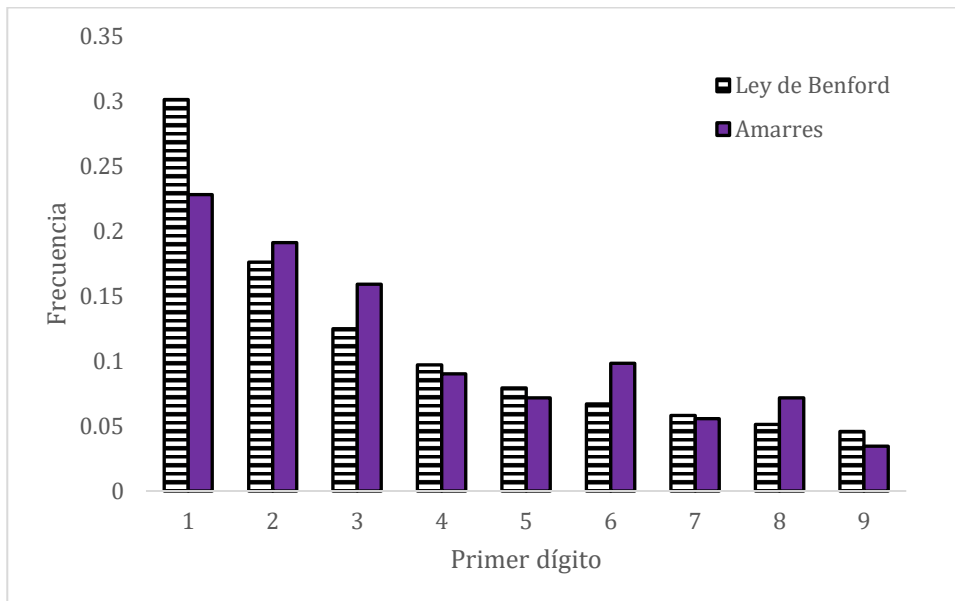


Figura 19. Proporciones del primer dígito para el número de amarres frente a la Ley de Benford.

En cuanto al número de habitantes, presenta un alto nivel de conformidad con un MAD de 0.001, pese al dígito 4 que se desvía del patrón establecido por Benford (Figura 20). En cambio, si el número de habitantes se expresa en densidad de población, es decir, en habitantes por m^2 , el nivel de conformidad es bajo con una MAD de 0.09. Pues bien, la densidad de población es una variable acotada que se distribuye en un único orden de magnitud haciendo que la probabilidad del primer dígito no siga ningún patrón (Figura 20).

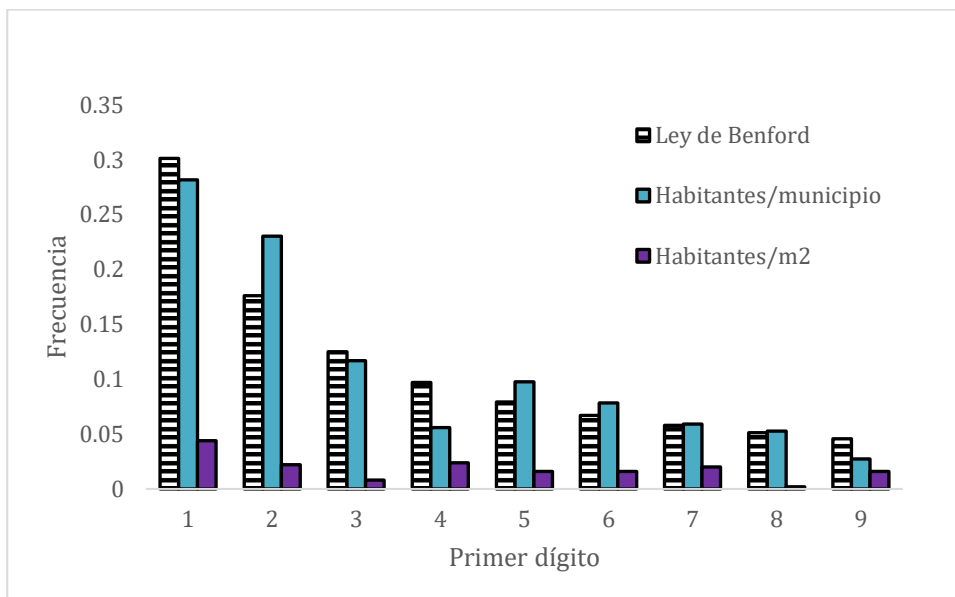


Figura 20. Proporciones del primer dígito para el número de habitantes por municipio y por metro cuadrado frente a la Ley de Benford

Se muestra a continuación una tabla resumen con los resultados de la prueba de bondad de ajuste y el nivel de conformidad para cada variable:

Tabla 4. Resumen de los resultados de la prueba de bondad de ajuste.

Variable	MAD	Nivel de conformidad
Longitud de playas	0.007	Alto
Erosión crónica	0.0093	Alto
Erosión eventual	0.007	Alto
Área de cuencas hidrográficas	0.016	Alto/Aceptable
Altura de ola	0.0426	Marginalmente aceptable
Periodo de pico	0.11	Bajo
Marea astronómica	0.085	Bajo
Marea meteorológica	0.0359	Marginalmente aceptable
Área de yacimientos subacuáticos	0.012	Alto
Longitud de las ARPSIs	0.011	Alto
Lámina inundable Tr100	0.0178	Aceptable
Lámina inundable Tr500	0.015	Alto
Habitantes/ZI Tr 100	0.0097	Alto
Habitantes/ZI Tr 500	0.0080	Alto
Habitantes/municipio	0.0011	Alto
Habitantes/m2	0.092	Bajo
Número de amarres	0.022	Aceptable

5. DISCUSIÓN

En este apartado, se interpretan los resultados en base a las condiciones de cumplimiento enumeradas en el Apartado 2. Con el fin de mostrar evidencias acerca de la relevancia de las distintas condiciones, se someten algunos conjuntos de datos a variaciones que favorezcan o evidencien la relevancia de las condiciones de cumplimiento.

La intervención humana en toma de datos para las distintas variables no se puede considerar un factor limitante para el cumplimiento de la Ley de Benford. Así pues, muchas de las variables son producto de mediciones bajo el criterio humano y aun así presentan un nivel de conformidad alto, como por ejemplo la longitud de las ARPSIs o el área de los yacimientos submarinos. A pesar de que los valores de estas variables no tengan una ocurrencia natural, tampoco han sido asignados de manera secuencial ni de manera aleatoria. Estos valores son establecidos siguiendo el mismo criterio para cada conjunto de datos, como pueden ser los criterios de vulnerabilidad para las ARPSIS o el volumen de arena para los yacimientos submarinos.

El tamaño de la muestra es una condición que en el estado del arte no está del todo definida. Es evidente que el número de datos condiciona el nivel de conformidad, pero no se define un umbral claro. Agregar los datos de playas y cuencas por provincias, pone en evidencia la estrecha relación entre el tamaño de la muestra y el nivel de conformidad ya que cada provincia tiene un número muy variable de datos. En la Figura 21 se representa la desviación media absoluta en

función del número de datos de la muestra de longitud de playas (izquierda) y área de las cuencas (derecha), donde podemos observar un mismo patrón, cuanto menor es el tamaño de la muestra, menor es el nivel de conformidad. En la longitud de playas podemos observar que a partir de los 100 datos el nivel de conformidad varía entre aceptable y alto. En el caso del área de las cuencas, las muestras por provincia son reducidas por lo que no llega a alcanzar un nivel de conformidad bueno, pero si se puede observar esa tendencia de disminución de la MAD con el aumento de datos. Es decir, para que una variable se ajuste a la distribución de Benford, la muestra debe superar los 100 datos.

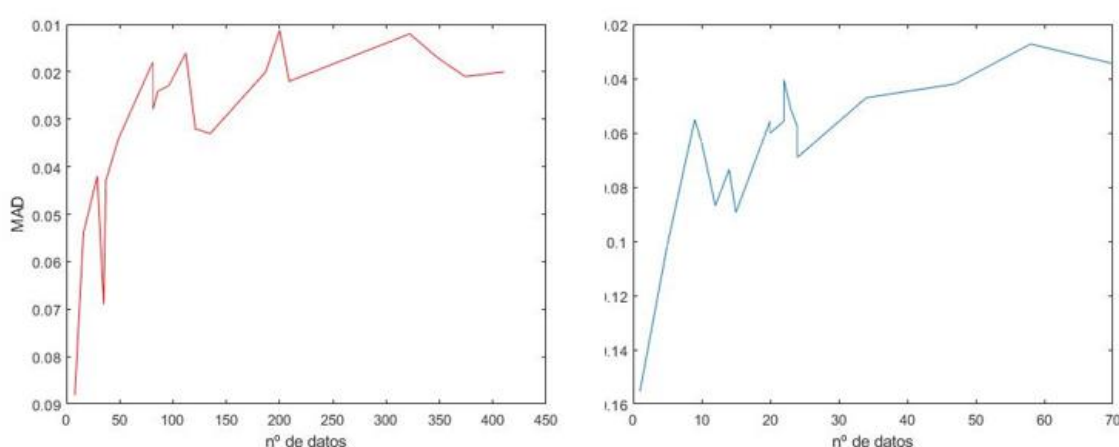


Figura 21. Distribución de MAD en función del número de datos de playas y cuencas agregados por provincias.

Si la agrupación de los datos de longitud de playas, en vez de realizarla por provincia, se realiza de manera aleatoria, la forma de la curva es semejante al de la figura anterior, pero el nivel de conformidad para cada tamaño de la muestra es superior como se observa en la Figura 22.

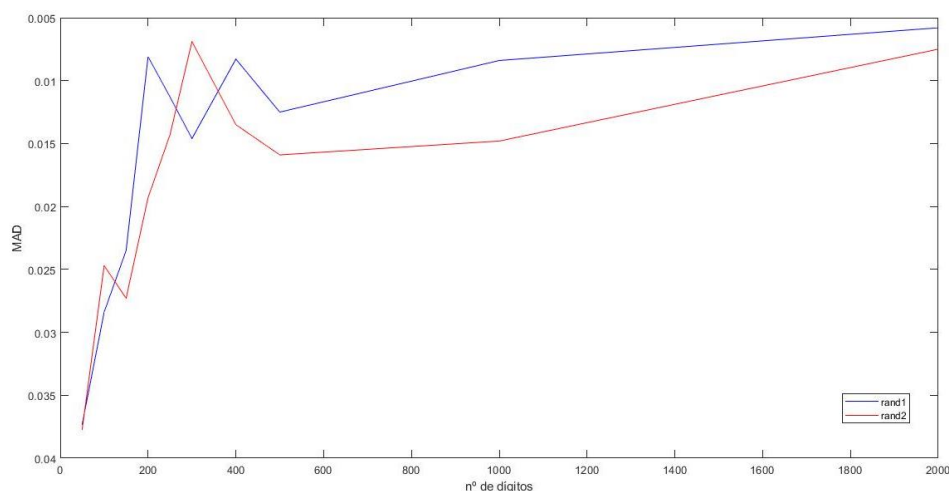


Figura 22. Distribución de MAD en función del número de datos para conjuntos aleatorios de playas.

Esta agregación por provincias revela una cierta coincidencia en MAD para cada muestra (Tabla 5), confirmando una estrecha relación entre el nivel de conformidad y la fuente de procedencia de los datos (personal técnico de cada provincia).

Tabla 5. Resultados de MAD de las variables de erosión para cada provincia.

Provincia	MAD	
	Erosión crónica	Erosión eventual
Melilla	0.04575749	0.04575749
Ceuta	0.15018805	0.15018805
Huelva	0.05199566	0.05063326
Bizkaia	-	0.06910015
Gipuzkoa	0.11467541	0.10953678
Granada	0.08133559	0.06029152
Lugo	0.05076574	0.04816931
Cantabria	0.02266553	0.03440191
Cádiz	0.04848699	0.03154202
Castellón	0.02579599	0.02579599
Barcelona	0.032	0.03
Almería	0.009	0.0408581
Málaga	0.01746477	0.01461331
Girona	0.06220303	0.04863466
Murcia	0.05	0.025
Asturias	0.02210026	0.02210026
Las Palmas	-	-
Baleares	0.01786688	0.06334358
Pontevedra	0.05	0.036
A Coruña	0.076	0.035

Otra condición de cumplimiento que se menciona es la presencia de varios órdenes de magnitud dentro de un conjunto de datos. El efecto de esta condición en los niveles de conformidad sí que se ve reflejado en los resultados. Los niveles de conformidad de variables con un único orden de magnitud son bajos. Las variables con un rango de valores acotado, altura de ola, periodo de pico, mareas meteorológica y astronómica y densidad de población (hab./m²), presentan un mal ajuste frente a la Ley de Benford. Sin embargo, no se pueden mostrar evidencias de que, a mayor rango de valores, mayor es el nivel de conformidad, ya que las variables no acotadas están sujetas a otros factores que condicionan su cumplimiento. Por ejemplo, el número de amarres y la longitud de playas constan de tres órdenes de magnitud entre el mínimo y el máximo, pero su MAD es muy diferente, 0.022 o 0.007 respectivamente.

Una peculiaridad de esta ley es que es invariante ante cambios de escala. Poniendo a prueba esta condición observamos que, efectivamente, la desviación media absoluta se mantiene igual modificando las unidades de medida. En un primer

momento, se planteó la posibilidad de mejorar el ajuste de aquellas variables redondeadas variando la unidad de medida, por ejemplo, si el área de las cuencas se expresa en Km, estaríamos eliminando esa fracción redondeada. Sin embargo, la distribución de probabilidad se mantiene igual, conservando el mismo resultado de MAD.

6. CONCLUSIÓN

El estudio de las frecuencias del primer dígito para los conjuntos de datos empleados para la gestión costera es muy dispar. Así pues, los resultados obtenidos para las distintas variables no muestran unanimidad frente al ajuste de la Ley de Benford. Esto es debido a las condiciones que limitan el cumplimiento de la ley. Las distintas características de las variables sometidas al estudio han permitido identificar los requisitos para que un conjunto de datos siga la distribución de frecuencias del primer dígito establecida por Benford. Con este primer hito se consigue descartar aquellas variables que no son aptas para el cumplimiento de dicha ley matemática. Pues bien, una de las características más restrictivas que limitan el cumplimiento son los valores acotados. En consonancia con Nigrini en 1997, se determina que las variables cuyo rango de valores es reducido no presenta una frecuencia del primer dígito predecible. Con ello, se descartan las variables de clima marítimo (altura de ola, periodo de pico, marea meteorológica y marea astronómica) y densidad de población (habitantes por m^2), todas ellas con un nivel de conformidad marginalmente aceptable o bajo.

Se ha verificado que el tamaño de la muestra es relevante para el ajuste. Se determina que para que una variable siga la distribución de Benford, ha de tener más de 100 datos para que su nivel de conformidad sea aceptable y más de 200 datos para que el nivel de conformidad sea alto.

Geyer y Martí en 2012 detectaron errores en algunos conjuntos de datos vulcanológicos debido al redondeo. Atendiendo a que el redondeo es la única limitación de variables como la longitud de playas y el área de las cuencas hidrográficas, se puede justificar con ello su ligera desviación frente a la Ley de Benford.

Las variables sujetas al criterio humano como las ARPSIs, los yacimientos submarinos, incluso el número de amarres en los puertos deportivos y el número de habitantes por municipio, demuestran con su alto nivel de conformidad que la ley es válida para aquellas variables que no son naturales. Pues bien, como se venía comentando a lo largo del estudio, la Ley de Benford se ve limitada si el conjunto de datos atiende a una asignación secuencial o a un proceso completamente aleatorio, de no ser así la distribución de los primeros dígitos es predecible según establece Benford, independientemente de si está influenciada por el pensamiento humano.

Un claro ejemplo de ello son los resultados de las consultas técnicas. Los valores de erosión, tanto crónica como eventual, presentan un nivel de conformidad alto. No

obstante, la coincidencia en los resultados de la desviación media absoluta para cada provincia revela un vínculo estrecho entre el nivel de conformidad con la Ley de Benford y la fuente de datos.

Tras comprobar si las distintas variables del ámbito costero se ajustan a la Ley de Benford y definir los límites de aplicación de la ley, se concluye que la Ley de Benford tiene potencial como una herramienta de control de calidad para algunas bases de datos. Sin embargo, es importante definir bien el rango de aplicabilidad de la misma. Si las variables a testar cumplen aquellas condiciones definidas en el presente estudio, la Ley de Benford puede ayudar a validar y analizar parte de la información incluida en bases de datos destinadas a la gestión de la costa.

REFERENCIAS

Deckert, J., Myagkov, M., & Ordeshook, P. C. (2011). Benford's Law and the detection of election fraud. *Political Analysis*, 19(3), 245-268.

Dumas, C. F., & Devine, J. H. (2000). *Detecting evidence of non-compliance in self-reported pollution emissions data: An application of Benford's Law* (No. 372-2016-19503).

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting*, 5(1), 17-34.

Fewster, R. M. (2009). A simple explanation of Benford's Law. *The American Statistician*, 63(1), 26-32.

Geyer, A., & Martí, J. (2012). Applying Benford's law to volcanology. *Geology*, 40(4), 327-330.

Hj, T. P. A Statistical Derivation of the Significant-Digit Law.

Morales, H. R. (2020). Análisis del funcionamiento de medidores de energía eléctrica a partir de aplicar la Ley de Benford. In *VI Simposio Argentino de Ciencia de Datos y GRANdes Datos (AGRANDA 2020)-JAIIO 49 (Modalidad virtual)*.

Nigrini, M. J., & Mittermaier, L. J. (1997). The use of Benford's law as an aid in analytical procedures. *Auditing*, 16(2), 52.

Nigrini, M. J., & Miller, S. J. (2007). Benford's law applied to hydrology data—results and relevance to other geophysical data. *Mathematical Geology*, 39(5), 469-490.

Nigrini, M. (2012). *Benford's Law Applications for Forensic Accounting, Auditing, and Fraud Detection*. Wiley.

Sambridge, M., Tkalčić, H., & Jackson, A. (2010). Benford's law in the natural sciences. *Geophysical research letters*, 37(22).

Tödter, K. H. (2009). Benford's Law as an Indicator of Fraud in Economics. *German Economic Review*, 10(3), 339-351.

Yunta, L. R. (1998). Evaluación e indicadores de calidad en bases de datos. *Revista española de documentación científica*, 21(1), 9-23.