02-037

MERI: METHODOLOGY FOR ESTIMATING COST AND TIME RISKS IN PROJECTS USING MACHINE LEARNING. APPLICATION TO OIL&GAS INFRASTRUCTURE.

Alonso Iglesias, Guillermo (1); Ortega Fernández, Francisco (1); Villanueva Balsera, Joaquín (1); Vergara González, Eliseo (1)

(1) Universidad de Oviedo

Historically, cost overruns in construction projects have been a major problem. Nowadays, the magnitude of this problem is completely unsustainable with cases of "billionaire" cost overruns all over the world. Previous works stresses the role of risk management as one of the best systems for understanding the key aspects of the project that give rise to these deviations, as well as controlling and mitigating them. The main objective of the work is the development of a methodology that allows the identification of those risks that most influence the cost overruns of a specific project portfolio, associating them to the different stages of the project, in order to make a preferential distribution of control resources. To achieve this objective, the methodology uses techniques based on Artificial Intelligence, specifically SOM modelling. Furthermore, in order to link the risks to the stages of the project, the methodology proposes a parameterisation of the costs throughout the life cycle of the project by means of a Beta distribution. Finally, as a validation of the methodology, a case study is carried out with Oil & Gas Offshore infrastructure projects.

Keywords: Cost Estimation; Project Risk Management; Oil&Gas; Machine Learning.

MERI: METODOLOGÍA DE ESTIMACIÓN DE RIESGOS DE COSTE Y PLAZO EN PROYECTOS MEDIANTE MACHINE LEARNING. APLICACIÓN A INFRAESTRUCTURA OIL&GAS.

Históricamente, los sobrecostes en los proyectos de construcción han sido uno de sus principales problemas. En la actualidad, la magnitud de esta problemática es completamente insostenible con casos de sobrecostes “milmillonarios” por todo el mundo. Los trabajos realizados hasta el momento para resolver el problema inciden en el papel de la gestión del riesgo como uno de los mejores sistemas para conocer los aspectos clave del proyecto que dan lugar a estas desviaciones, así como controlarlas y mitigarlas. El objetivo principal del trabajo es el desarrollo de una metodología que permita identificar aquellos riesgos que más influyen en los sobrecostes de una cartera de proyectos concreta, asociándolos a las distintas etapas del proyecto, con el fin de realizar una distribución preferencia de los recursos de control. Para conseguir este objetivo, la metodología utiliza técnicas basadas en Inteligencia Artificial, específicamente la modelización SOM. Además, para vincular los riegos a las etapas del proyecto, se propone en la metodología una parametrización de los costes a lo largo del ciclo de vida del proyecto mediante una distribución Beta. Por último, como validación de la metodología, se realiza un caso de estudio con proyectos de infraestructura Oil & Gas Offshore.

Palabras clave: Estimación de costes; Gestión de riesgos; Oil&Gas; Machine Learning

Correspondencia: fran@api.uniovi.es

Agradecimientos: Este trabajo ha sido subvencionado a través del programa de “Ayudas para Grupos de Investigación de Organismos del Principado de Asturias” (GRUPIN 2021-2023) de la Fundación para el Fomento de Asturias de la Investigación Científica Aplicada y la Tecnología (FICYT) -Gobierno del Principado de Asturias, (Ref: SV-PA-21-AYUD-2021-50953). Proyecto financiado por la Unión Europea a través del Fondo Europeo de Desarrollo Regional.

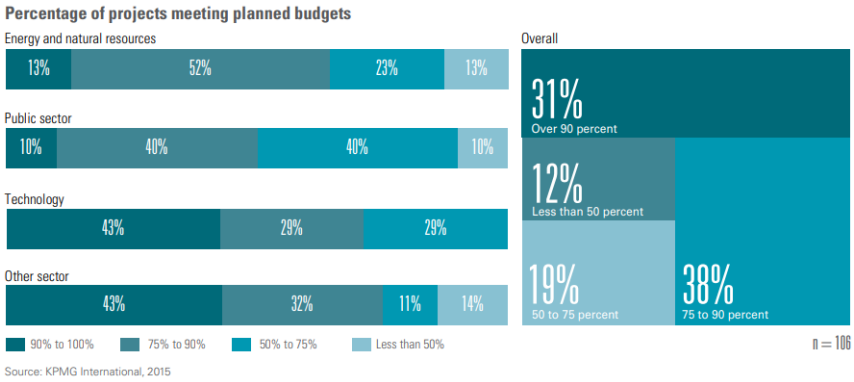
Introducción

El sector de la construcción, en todo su ámbito de competencias, se ha enfrentado históricamente a diversos problemas, pero seguramente el más persistente e invariante, ha sido el de los denominados sobrecostes o desviaciones.

El término "sobrecoste" no aparece en el diccionario de la Real Academia Española, pero diversos autores como Flyvbjerg et al., (2002) u Odeck, (2004) coinciden en definir este concepto como *'la diferencia entre los costes de construcción previstos en el momento de la toma de decisión de la construcción, y los costes reales incurridos a la finalización del proyecto'* (Gifra Bassó, 2018). Estas desviaciones aparecen cuando existe un imprevisto (es decir, un evento que no está contemplado en la planificación) que provoca un desajuste entre la previsión económica realizada y el coste final de las operaciones. Indudablemente, los desajustes y desviaciones en los proyectos no solo son presupuestarios. En algunos casos, las obras sufren también retrasos sobre el plazo acordado y ambas desviaciones están fuertemente vinculadas.

Como ejemplo para poder entender la magnitud de este problema, la consultora *KPMG*, una de las consultoras con más prestigio e influencia a nivel mundial, realizó un estudio en 2015 sobre el éxito en los proyectos (*Global Construction Survey 2015*, 2019). Para ello, se basó en una encuesta realizada a más de 100 directores de proyecto senior, tanto del ámbito público como privado. En los resultados de esta encuesta se observa que sólo el 31% de los proyectos de construcción finalizaron con una desviación final menor del 10% respecto al presupuesto original (Figura 1).

Figura 1: Porcentaje de proyectos en función de haber conseguido terminar dentro del presupuesto planeado



Estas desviaciones tienen un impacto global, ya que los sobrecostes en los proyectos afectan directamente a la economía y a los fondos de las organizaciones promotoras que, en el sector de la construcción, son generalmente instituciones públicas. Los sobrecostes obligan a destinar recursos extraordinarios que se sustraen de otras necesidades de la sociedad.

Este problema está ligado, por tanto, a los principios de sostenibilidad de la sociedad y su búsqueda mediante la innovación, cuya máxima expresión pueden ser los Objetivos de Desarrollo Sostenible (ODS) y muy especialmente, a dos: ODS 9 Industria, Innovación e Infraestructura y ODS 11 Ciudades y Comunidades Sostenibles.

Las organizaciones y consultoras líderes están invirtiendo en tecnología, recursos humanos y cultura de proyectos para estar más preparadas y combatir el problema de las desviaciones

en coste y plazo de los proyectos. Así, el estudio de 2019 realizado por *KPMG* (*Global Construction Survey 2019 - KPMG Global*, 2019) investiga la preparación de las organizaciones constructoras para afrontar el futuro. Para ello, se basa en un índice que engloba tres campos: Dirección y Control, Tecnología e Innovación y Recursos Humanos. Los mejores resultados se obtuvieron en empresas que constantemente están buscando oportunidades para mejorar y anticiparse a los riesgos y los cambios, con el objetivo de crear una mejor industria de la construcción para todas las partes interesadas.

No obstante, puede llegar a darse la paradoja de que, para reducir los posibles sobrecostes incurridos en una tarea, se realicen acciones de control y monitorización con mayor coste que la posible desviación. Además, es habitual que exista limitación de medios o recursos existentes para poder controlar y monitorizar el proyecto. Por tanto, para resolver esta problemática, conocer aquellos aspectos en los que, destinando el mismo número de recursos, se reduzca en mayor medida el posible sobrecoste, permitirá optimizar el resultado final del proyecto.

Una de las mejores herramientas para poder conocer estos aspectos clave que controlar y a los que destinar recursos es la gestión de riesgos en los proyectos. Por todo ello, el **objetivo principal** de este trabajo es **desarrollar una metodología que permita identificar aquellos riesgos que más influyen en los sobrecostes de una cartera de proyectos**, mediante el uso de técnicas de modelización y minería de datos, sobre una base de datos de casos históricos.

Así surge la metodología **MERI** (**Me**todología de **R**iesgos basada en **I**nteligencia Artificial), que será desarrollada en esta comunicación. Además, se validará y discutirá mediante su aplicación a un caso de estudio, utilizando la base de datos de una investigación externa, que aportará veracidad y objetividad a los resultados.

Dicho caso de estudio utiliza una base de datos fruto de la monitorización de 15 proyectos de construcción de infraestructuras *Oil&Gas Offshore*, en los que cada mes, se recogía que riesgo (de 12 previamente identificados) había sido el más importante y en cuánta desviación porcentual se había incurrido dicho mes.

Herramientas utilizadas en la metodología

En el desarrollo y aplicación de la metodología se han de utilizar principalmente dos herramientas. La primera de ellas se trata de la tipología de modelo de Inteligencia Artificial que se usará para relacionar y cuantificar los riesgos y los sobrecostes del proyecto. Para el desarrollo e implementación de dicho modelo, se propone seguir la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) para asegurar obtener un modelo válido, mediante técnicas de minería de datos. La segunda herramienta en la que se apoya MERI, es la parametrización de los costes en el ciclo de vida del proyecto, mediante una distribución Beta. Su aplicación permite transformar valores de desviaciones relativos o categóricos, a valores absolutos teniendo en cuenta la distribución de los costes a lo largo del proyecto.

2.1 Jerarquización de riesgos: Modelización SOM

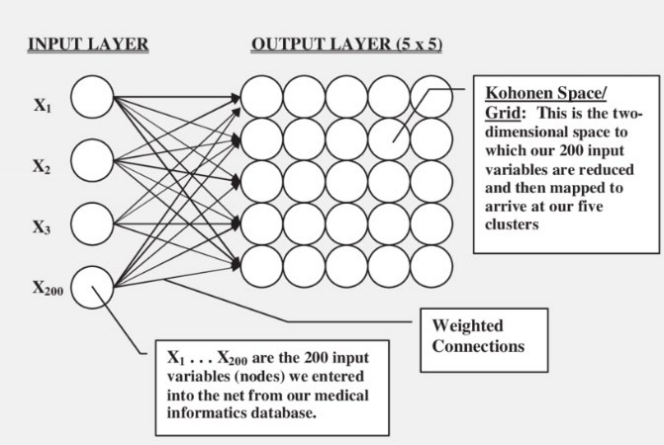
Para categorizar los proyectos, se utiliza el modelo de tipo *clustering* SOM *(Self-Organized Maps*). La aplicación de esta tipología de técnica permite analizar, buscar similitudes y jerarquizar los riesgos de una base de datos, en función de variables como la desviación recogida o la variable temporal. Dicha aplicación, aporta una novedad al estado del arte actual en el análisis de la gestión de riesgos, dónde los modelos de tipo *clustering* no han sido aún explotados. No obstante, dependiendo del caso concreto de estudio, la metodología se podría adaptar a otro tipo de modelos de Inteligencia Artificial.

Un Mapa Auto-Organizado (SOM), es una clase de Red Neural Artificial (RNA) no supervisada presentada en 1982 por T. Kohonen. Este modelo se basa en ciertas evidencias descubiertas a nivel cerebral, y realiza una reducción de la dimensionalidad del espacio de entrada para

producir mapas ordenados topológicamente. Este tipo de red posee un aprendizaje no supervisado competitivo. La propia red es la encargada de auto-organizarse y descubrir rasgos comunes, regularidades, correlaciones o categorías en los datos de entrada (Wehrens & Buydens, 2007).

El algoritmo SOM procesa cada registro de entrada normalizado y permite el proceso incluso ante la presencia de valores nulos en los datos de entrada (Kohonen, 2001). El uso de este algoritmo comienza con una etapa de aprendizaje cuyo objetivo es categorizar los datos que se introducen en la red. Los valores similares deben clasificarse en la misma categoría y, por tanto, deben activar la misma neurona de salida. Al tratarse de un método no supervisado, las clases o categorías deben ser creadas por la propia red a través de las correlaciones entre los datos de entrada.

Figura 2: Ejemplo de mapa auto-organizado de Kohonen (Castellani & Castellani, 2003).



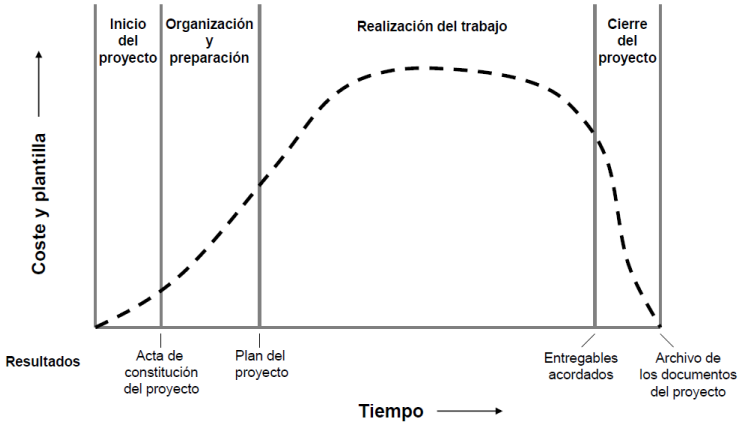
Además, SOM también puede aplicarse al reconocimiento de patrones (aprendizaje supervisado). La información se da al final de entrenamiento: si se trata de regresión, la estimación es la media de todos los elementos en el nodo; si es de clasificación, se usa la estrategia *winner-takes-all*. Otra opción es utilizar la información de salida para el entrenamiento, teniendo en cuenta para este tanto las distancias de las características al nodo como las de las salidas. Se generan dos mapas (capas), uno para los atributos y otro para las salidas. Este principio puede ser extendido a más capas, generando mapas super-organizados. Para cada capa se calcula un nivel de similitud y las similitudes individuales se combinan en un solo valor que se usa para determinar el nodo ganador.

2.2 Parametrización del ciclo de vida – Distribución Beta

La segunda componente del método es la parametrización de los costes a lo largo del ciclo de vida del proyecto. El objetivo de parametrizar el coste del proyecto en función tiempo, es poder representar la variabilidad de los costes totales de todas las tareas que se están realizando en cada instante del proyecto.

En el ámbito de la construcción, y parte del industrial, la tipología de ciclo de vida del proyecto mayoritaria es el tipo predictivo. Este tipo de ciclo de vida suele estar dividido en varias fases, habitualmente correspondientes a: Inicio del proyecto, Organización y preparación, Realización del trabajo, Cierre del proyecto (Project Management Institute, 2013). La mayor parte de los recursos son destinados a las partes centrales del proyecto, con lo que las desviaciones en términos porcentuales son mayores en términos absolutos en comparación con las otras fases del proyecto. De esta forma, parametrizando los costes a lo largo del proyecto, se consigue por tanto transformar esas desviaciones porcentuales a desviaciones absolutas comparables a lo largo de todo el proyecto.

Figura 3: Evolución de costes y plantilla en un ciclo de vida predictivo (Project Management Institute, 2013)



En el campo de la gestión de proyectos, históricamente se han utilizado diferentes distribuciones de probabilidad para modelizar comportamientos e incertidumbres en los proyectos, principalmente los costes y las duraciones de las tareas del proyecto. Las más utilizadas han sido la distribución triangular (Johnson, 1997), la distribución trapezoidal (Garvey et al., 2016), la distribución uniforme (Romero López, 2010), la distribución normal y log-normal (Schwarz & Sánchez, 2015), o la distribución Parkinson (Trietsch et al., 2012).

Pero, sin duda, la más populares y utilizada en los procesos estocásticos, ha sido la distribución Beta (Hahn & López Martín, 2015). Esto se debe a sus características intrínsecas que permiten que pueda ser muy adaptable en función de sus parámetros definitorios. De esta forma, con la variación de dichos parámetros, la distribución puede adoptar una amplia variedad de formas, con distintas intensidades en su asimetría y curtosis.

Esta característica es clave en casos como el que atañe esta parametrización del ciclo de vida, donde se comprueba que, en los proyectos constructivos, los costes o recursos se concentran en las zonas intermedias-finales.

Con la selección de los parámetros concretos (adaptados al caso de estudio), es posible representar la distribución de costes del proyecto en función del tiempo mediante esta distribución. Los valores mínimos y máximos serán los costes del primer y último mes del proyecto. El valor de la mediana de la distribución coincidirá con el mes de mayor gasto del proyecto. De esta forma, un experto en la gestión de proyectos puede fácilmente parametrizar el coste del proyecto a lo largo del ciclo de vida del proyecto, de una forma muy intuitiva.

3. MERI: Metodología de Análisis de Riegos basada en Inteligencia Artificial. *Oil&Gas case study*.

En este apartado, se expone la metodología objeto de esta investigación. A partir de los datos históricos de unos proyectos de la misma cartera (en la que se recogen los riesgos y desviaciones causadas por los mismos), se jerarquizarán los riesgos identificados en función de su influencia en incurrir en mayores desviaciones. Es decir, se seleccionarán aquellos riesgos más importantes y por tanto a los que destinar recursos para su control, ya que de esta forma se podrá reducir más eficientemente la desviación final del proyecto.

Para una mejor comprensión de la metodología, se acompaña la explicación de un caso de estudio real, cuya base de datos se ha obtenido de un estudio externo. La investigación de la que se ha obtenido la base de datos es: *An Approach Based on Bayesian Network for Improving Project Management Maturity: An Application to Reduce Cost Overrun Risks in Engineering Projects* (Sanchez et al., 2020).

En dicho trabajo, se desarrolló el marco general y un método para estimar el impacto que tiene la madurez (nivel de desarrollo) en gestión de proyectos (*Project Management Maturity*).

En este estudio, para ilustrar la aplicación del método y comprobar que se reduce el riesgo de que se supere el coste estimado del proyecto, se emplearon 15 proyectos *oil & gas offshore.* Los principales riesgos de estos proyectos (denominados *drift factors* en el estudio), así como las desviaciones causadas, conforman una base de datos perfecta para la aplicación de la metodología MERI como caso de estudio.

3.1 Definición del contexto de análisis

El primer paso para desarrollar la metodología es definir el campo de estudio y aplicación de la misma. Se ha de identificar la cartera de proyectos sobre la que el resultado será aplicable, acotando las similitudes que son necesarias mantener tanto en los proyectos que conforman la base de datos del modelo, como en los futuros proyectos en los que se podrá aplicar.

Se han de tener en cuenta características como el tamaño de los proyectos, los agentes involucrados, las regiones dónde se van a llevar a cabo, el contexto socioeconómico local y global, el marco legislativo al que se someten, u otras características identificativas de los proyectos objeto de la metodología, que puedan limitar la comparación entre los utilizados para modelizar y en los que se aplicará finalmente.

La investigación de Sánchez, F. et al. (2020) se centró en el caso de los proyectos de construcción de infraestructura *oil & gas offshore*. Por tanto, el contexto del caso de estudio se limita a esta tipología de proyectos.

Los datos de los 15 proyectos fueron recogidos entre los años 2013 y 2017. Cada uno ellos realizó un informe en el que se reflejaron las causas principales de las desviaciones sufridas durante su ejecución. Los detalles que se adjuntaron ha sido información como la fecha, la cantidad del sobrecoste, los factores y acciones para llevar corregir estas desviaciones, entre más información. Las características de estos proyectos son perfectamente comparables debido a que se llevaron a cabo en un periodo temporal similar y en regiones similares, evitando condicionantes externos diferentes.

3.2 Identificación de riesgos

Una de las opciones para realizar la identificación de riesgos, es la aplicación de las técnicas de IA como puede ser el *Case-Based Reasoning*, entre otras técnicas.

La otra opción es utilizar técnicas cualitativas de identificación, recogidas la mayoría de ellas en las metodologías de gestión de riesgos como el PMBoK, PRINCE2, etc. En este abanico se encuentran técnicas como:

Diagrama de Ishikawa

Listas de comprobación (*Checklists*)

Análisis de la Estructura de Descomposición del Proyecto (EDP, *WBS Analysis* en inglés)

Técnicas de *brainstorming*, o reuniones de expertos como el método Delphi

Diagramas de flujo casusa-efecto

Diagramas *Mind Mapping*

Entrevistas a expertos en el campo concreto de la cartera de proyectos a estudiar

En el caso de estudio de este trabajo, se ha utilizado la técnica de juicio de expertos para la identificación de los riesgos más relevantes. Se consultó a seis expertos en dirección de proyectos (dos consultores senior, dos directores de proyecto, y dos *partners* de una empresa consultora de gran tamaño) que satisfacían los siguientes requisitos:

Más de diez años de experiencia en dirección y gestión de proyectos (PM).

Haber participado en al menos dos de los quince proyectos que conforman la base de datos.

Conocimientos acreditados en dirección y gestión de proyectos mediante certificación del PMI (*Project Management Institute*).

Experiencia en resolución de problemas en proyectos con desviaciones.

Los expertos identifican en total 12 riesgos, después de varios filtrados y comparaciones. Para validar el conocimiento de los expertos, y, por tanto, la identificación de los riesgos principales, se realizó una búsqueda bibliográfica en la literatura de dirección y gestión de proyectos. Con esta investigación, se buscó comprobar que, en proyectos similares, otros estudios identificaron los mismos riesgos y factores de desviación. A continuación, se muestra en la Tabla 1, los riesgos identificados con las diferentes referencias de otras investigaciones contrastadas:

Tabla 1: Riesgos identificados para infraestructura off-shore (Sanchez et al., 2020)

ID	Riesgo / Factor de desviación
D1	No asignar el director de proyecto correcto. Falla en involucrar al equipo en el proyecto.
D2	Falta de comunicación. El equipo trabaja en silos.
D3	Requerimientos y cuestiones de política contractual
D4	Ausencia de análisis de riesgos y estimación de contingencias
D5	Falta de especificación en los requisitos del alcance
D6	Ausencia de una métrica para monitorizar desviaciones
D7	Los informes no reflejan la realidad
D8	Overoptimistic bias (tiempo y coste)
D9	Cuestiones de calidad
D10	Ausencia de un sistema que sigue los cambios a partir de datos históricos
D11	Retrasos en los aprovisionamientos y subcontrataciones
D12	Falta de integración horizontal

3.3 Creación de la base de datos

La modelización de esta metodología está basada en el uso de un conjunto de datos históricos, sobre los que se aplicará la técnica de modelización SOM. Por tanto, el próximo paso es la creación de la base de datos del modelo.

Mediante la monitorización de varios proyectos representativos, se obtendrá la información de las variables a modelar posteriormente. Cada caso particular puede seleccionar el tipo y cantidad de variables que se requieran, gracias a la capacidad de las técnicas SOM de modelar tanto con variables categóricas como variables continuas.

Las variables que nunca deberían faltar en la monitorización son la percepción, actuación o influencia de los riesgos, la variable temporal, y las desviaciones recogidas en cada uno de los períodos temporales considerados.

Sin embargo, en muchas de las ocasiones, no es fácil ni directo obtener estos valores de desviaciones en coste sobre el proyecto original. Las causas principales pueden ser, por ejemplo, identificar a qué tarea en concreto corresponden las desviaciones dentro de todas las tareas realizadas en ese período, la falta de trazabilidad en el seguimiento del proyecto, y por supuesto, la falta de medios o recursos, entre otras causas. Por ello, una opción plausible es categorizar las desviaciones dentro de unos rangos de porcentajes preestablecidos.

La base de datos del casos de estudio (Sánchez et al., 2020) consiste en la recopilación mensual de las desviaciones de 15 proyectos, durante los 4 años de su ejecución. Para cada mes de cada proyecto se identificó el riesgo/factor más importante y determinante, junto al rango de desviación en el coste que se produjo en dicho mes respecto al coste planificado. Esto da lugar a un total de 720 eventos, en los que se identifica un riesgo/factor y la desviación producida.

La desviación recogida en cada mes, respecto a la planificación de trabajos de dicho mes, se ha recogido en formato de rangos, no como una variable continua. Se seleccionaron cuatro rangos de desviación, con relación de escala logarítmica en base 10.

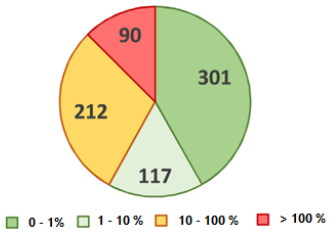
1. Sobrecoste < 1%: Sobrecoste incurrido para el mes de estudio, inferior al 1% del planificado para dicho mes. Considerado como riesgo y desviación aceptable. A la hora de utilizar este rango dentro de las modelizaciones SOM, se ha otorgado un valor numérico de 0.

   1% < Sobrecoste < 10%: Sobrecoste incurrido para el mes de estudio entre el 1 y 10 %, del planificado para dicho mes. Considerado como riesgo y desviación aceptable. A la hora de utilizar este rango dentro de las modelizaciones SOM, se ha otorgado un valor numérico de 0.5, conservando la escala logarítmica.

10% < Sobrecoste < 100%: Sobrecoste incurrido para el mes de estudio entre el 1 y 10 % (del planificado para dicho mes). Considerado como riesgo y desviación indeseable. A la hora de utilizar este rango dentro de las modelizaciones SOM, se ha otorgado un valor numérico de 1.5, conservando la escala logarítmica.

100% < Sobrecoste: Sobrecoste incurrido para el mes de estudio superior al 100% del planificado para dicho mes. Considerado como riesgo y desviación inaceptable. A la hora de utilizar este rango dentro de las modelizaciones SOM, se ha otorgado un valor numérico de 2.5, conservando la escala logarítmica.

Figura 4: Distribución de las desviaciones categóricas de los eventos en la base de datos (Sánchez et al., 2020)



En el estudio que recoge esta base de datos no se da importancia a la variable temporal dentro de su modelización. Así, no se tiene en cuenta la influencia del momento específico en el que se produce la desviación/riesgo dentro del ciclo de vida del proyecto. Para tenerla en cuenta, se ha utilizado una parametrización de la distribución de recursos a lo largo del proyecto mediante una función de tipo Beta, obteniéndose así una nueva variable denominada “Desviación ajustada”.

3.4 Estudio y preparación de los datos

Previo a desarrollar el modelo en sí, es necesario analizar las características de la base de datos del proyecto. De este análisis, se van a poder extraer conclusiones como la necesidad o no de realizar un pretratamiento de algún subconjunto de datos, detectar posibles incongruencias y *outliers* de la base de datos, así como identificar las limitaciones de la base de datos. En el caso de esta metodología, el foco de atención del análisis debería centrarse en:

La relación/comparación entre los riesgos y los proyectos

Los riesgos en relación con los meses del proyecto

Estadística descriptiva de las desviaciones para cada riesgo

En cada proyecto, representar las relaciones entre riesgos y desviaciones

Estadística descriptica de las desviaciones por cada proyecto

Relación entre las desviaciones y los meses

Para el global de la base de datos, estadísticas de las desviaciones

Si se recogen desviaciones categóricas y ajustadas, analizar su interrelación

Para estudiar estas relaciones, se pueden utilizar tanto tablas, como representaciones gráficas en 2D, e incluso con más de dos variables representadas.

Por ejemplo, en el caso de estudio de esta investigación, en la Tabla 2 se recogen las apariciones totales de cada riesgo en el global de la base de datos. Los casos D12 y D9 están poco representados, mientras que el riesgo D3 es el más recogido con diferencia. O por ejemplo también, en la Tabla 3, se recoge la estadística descriptiva de las desviaciones ajustadas por proyecto, omitiendo los proyectos 11, 12, 13, 14 y 15, que no registraron desviaciones.

Tabla 2: Apariciones totales de los riesgos en la base de datos.

Riesgo	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
Recuento	24	34	125	26	63	49	21	26	14	26	19	9

Tabla 3: Estadística descriptiva de las desviaciones ajustadas por proyecto.

Proyecto	Q1	Mediana	Q3	Media	Desv. típica
1	0.57	1.41	4.92	2.55	2.16
2	1.29	2.03	3.92	2.51	1.77
3	1.05	1.74	4.11	2.41	1.92
4	1.00	1.89	4.17	2.47	1.84
5	0.46	1.58	3.65	2.15	2.13
6	1.09	1.60	4.92	2.64	1.92
7	1.20	3.19	5.50	3.53	2.84
8	1.34	4.02	5.03	3.37	2.12
9	0.98	4.33	5.50	3.81	2.90
10	0.00	0.00	0.00	0.12	0.38

Se observa claramente como los proyectos 7, 8 y 9 son los que han recogido una mayor desviación tanto ajustada, como mayor número de casos con mayor desviación (>100%). Mientras los seis primeros proyectos, las desviaciones se concentran en los dos rangos intermedios, en los tres mencionados con anterioridad, se concentran en los dos rangos con más desviación.

3.5 Modelización y evaluación

La modelización es la fase central de las metodologías, como el CRISP-DM, pero no se ha de olvidar su carácter iterativo, vinculada a la evaluación de resultados y el pretratamiento de datos a la hora de realizar un nuevo modelo. La primera decisión al realizar una modelización es la selección de la técnica a utilizar. En el caso de esta metodología, se ha seleccionado una técnica de *clustering* basada en RNA, denominada SOM *(Self-Organized Maps*).

En un entrenamiento no supervisado, el modelo genera un mapa agrupando las variables introducidas sin ningún conocimiento a priori, sin buscar relaciones entre ellas. Por otro lado, es posible realizar un entrenamiento supervisado. En este caso, el modelo conoce la variable que debe organizar el mapa, y busca relaciones entre el resto de las variables y esta, para crear el mapa resultado del modelo.

Sin embargo, una vez el mapa está realizado, es posible seleccionar los casos concretos que han sido agrupados en cada neurona del mapa. Esta posibilidad, da opción a realizar "proyecciones" de otras variables en el mapa resultado del entrenamiento. Se escogen los casos que han sido agrupados en cada neurona, y se cruzan con la base de datos del proyecto.

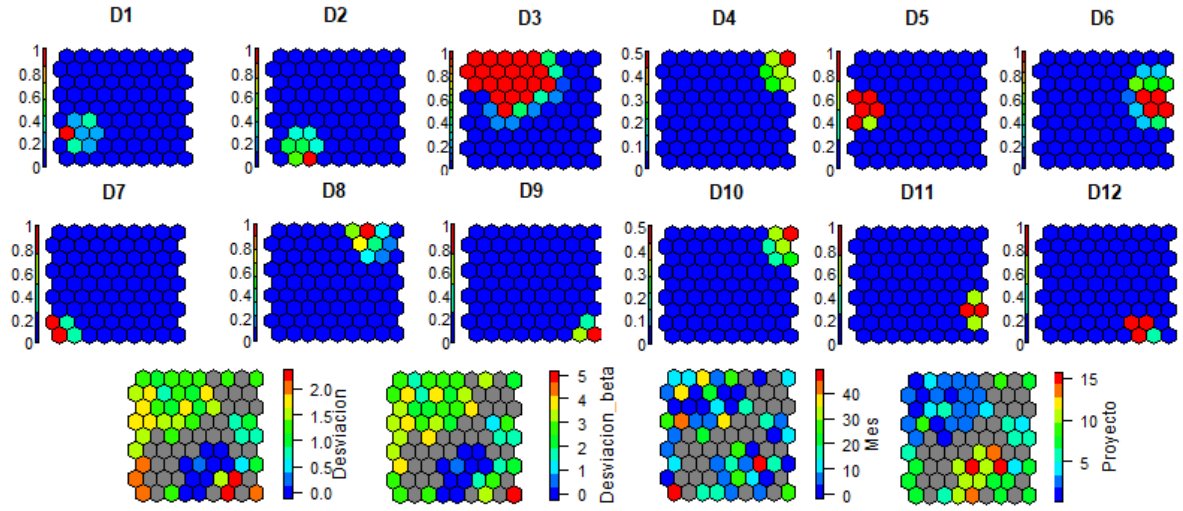
En esta metodología de gestión de riesgos, se recomienda comenzar por entrenar un modelo no supervisado, solamente con los riesgos identificados. Una vez generado el mapa, se proyectarán variables como los proyectos, los meses, y, por supuesto, las desviaciones (categóricas, ajustadas o reales). Comparando estos mapas, ya se pueden ir extrayendo conclusiones de la relación entre los riesgos y las demás variables.

Sobre la modelización y evaluación del caso de estudio, se realizó el entrenamiento no supervisado, en el que los inputs del modelo han sido solamente los 12 riesgos identificados. Obtenido el mapa, se "proyectaron" el resto de las variables en él, creando los mapas para las variables no entrenadas, tal y como se observa en la Figura 5.

En los mapas de componentes resultados del entrenamiento, se observa cómo se han *clusterizado* perfectamente cada uno de los riesgos, siendo muy pocas las neuronas que comparten casos con riesgos diferentes (aparecen con valores en mapas de riesgos distintos). En cuanto a las variables no entrenadas, es visible la clara relación entre las desviaciones consideradas categóricas (*Desviación*) y las ajustadas con la Beta (*Desviacion_beta*). Los casos de nula o baja desviación coinciden perfectamente, así como los casos de desviaciones medias y bajas de la parte superior izquierda del mapa.

En el caso de las desviaciones altas, sería necesario también fijarse en el mapa *Mes*. Para aquellos casos de desviación categórica elevada, y meses de proyecto intermedios (inferior derecha), se obtiene una desviación ajustada elevada.

Figura 5: Resultados del entrenamiento no supervisado del caso de estudio



Sin embargo, los casos de la esquina inferior izquierda corresponden con desviaciones categóricas elevadas y meses iniciales y finales del proyecto, con lo que la desviación

ajustada se corrige a la baja. A continuación, se analizan los riesgos considerados más relevantes, ordenados, a la hora de monitorizar y reducir desviaciones de acuerdo con los resultados:

D9 – Cuestiones de calidad: Se dispone en la esquina inferior izquierda, correspondiente con las desviaciones ajustadas más elevadas, ya que coinciden meses del proyecto intermedios y desviaciones categóricas elevadas.

D7 – Los informes no reflejan la realidad: El clúster se encuentra en la esquina inferior izquierda, con una desviación categórica elevada. Sin embargo, se presenta en meses iniciales y finales del proyecto, con lo que se rebaja su desviación mediante el ajuste con la Beta.

D5 – Falta de especificación en los requisitos del alcance: Muy localizado en la parte central izquierda, no comparte neuronas con los casos vecinos. Se le relacionan meses iniciales y finales, con lo que la desviación media-alta registrada por las desviaciones categóricas, se relativiza ajustándola con la Beta.

D3 – Requerimientos y cuestiones de política contractual: Es el grupo más numeroso y repartido de todos los riesgos, a causa de tener mucho mayor número de casos. Coincide con los casos de desviaciones medias en ambos casos, y con los proyectos 1-5.

Sin embargo, no menos importante, es también identificar aquellos riesgos que no merece tanto la pena monitorizar, por incurrir en bajas desviaciones:

D2 – Falta de comunicación: El equipo trabaja en silos: Se identifica con desviaciones medias-bajas, tanto para las desviaciones ajustadas, como para las desviaciones categóricas.

D11 – Retrasos en los aprovisionamientos y subcontrataciones: Este riesgo se localiza en un clúster en el que se recogen casos de desviaciones bajas o muy bajas, para meses iniciales-medios del proyecto.

5. Conclusiones y líneas futuras

Los sobrecostes en los proyectos de construcción han sido y son una, por no decir la mayor, problemática del sector de la construcción. Sus causas pueden ser múltiples, pero está claramente aceptado que la gestión de los riesgos es una de sus principales causas, así como una posible solución.

Durante este trabajo se ha desarrollado una nueva metodología que busca identificar y jerarquizar los riesgos de una cartera de proyectos, basándose en la modelización por técnicas de *clustering* de una base de datos de proyectos ya realizados. De esta investigación, se pueden obtener las siguientes conclusiones:

Es posible identificar y jerarquizar los riesgos de una cartera de proyectos, mediante el uso de técnicas de minería de datos y modelizaciones basadas en redes neuronales artificiales de clusterización.

La parametrización de los costes en las fases del ciclo de vida de un proyecto es posible mediante el uso de distribuciones Beta.

En el caso de estudio de este trabajo, se comprueba que la metodología es capaz de trabajar con variables de desviación categóricas y numéricas.

La metodología desarrollada en esta investigación ha sido validada mediante su aplicación en un caso de estudio para proyectos *Oil & Gas offshore*.

En la cartera de proyectos *Oil & Gas offshore*, los riesgos relacionados con el control de calidad y el alcance, son los más relevantes e importantes, incurriendo en mayores sobrecostes. De igual manera, los riesgos con menos influencia se identifican con la falta de comunicación y los relacionados con los aprovisionamientos.

Es difícil estimar la reducción de los sobrecostes que se puede alcanzar en un proyecto con esta metodología, pero se puede afirmar que los recursos serán utilizados más eficientemente, y se conseguirá una reducción de los sobrecostes en mayor o menor medida.

Aunque el trabajo de investigación se considera suficientemente válido para sostener las conclusiones anteriormente citadas y considerar como logrados los objetivos de esta, existen posibles mejoras a tener en cuenta, así como líneas futuras de investigación fruto de este trabajo:

Validar la metodología mediante más casos de estudios, de diferentes carteras de proyectos vinculadas a la construcción, tales como ingeniería ambiental, hidráulica, transportes, etc. Además, ampliar la modelización del caso de estudio actual a modelizaciones supervisadas para contrastar resultados.

Añadir a la metodología un proceso que permita integrar un análisis de riesgos durante la ejecución del proyecto.

Utilizar la metodología en un caso de estudio que contemple combinaciones de riesgos (afectando al mismo tiempo), para comprobar su validez a la hora de determinar las combinaciones de riesgos más perjudiciales en términos de sobrecoste del proyecto.

Mediante la aplicación en esta metodología en casos reales, cuantificar la eficacia de la misma con indicadores como el número de proyectos que termina con una desviación menor del 10%, etc.

6. Referencias

Castellani, B., & Castellani, J. (2003). Data Mining: Qualitative Analysis with Health Informatics Data. *Qualitative health research*, *13*, 1005-1018.

Flyvbjerg, B., Holm, M. S., & Buhl, S. (2002). Underestimating Costs in Public Works Projects: Error or Lie? *Journal of the American Planning Association*, *68*(3), 279-295.

Garvey, P. R., Book, S. A., & Covert, R. P. (2016). *Probability methods for cost uncertainty analysis: A systems engineering perspective*.

Gifra Bassó, E. (2018). *Desarrollo de un modelo para el seguimiento y control económico y temporal durante la fase de ejecución en la obra pública.* Tesis doctoral, Universidad de Girona.

Global Construction Survey 2015: Climbing the curve - KPMG Global. (2019, abril 2). KPMG. https://home.kpmg/xx/en/home/insights/2015/03/global-construction-survey.html

Global Construction Survey 2019—KPMG Global. (2019, mayo 1). KPMG. https://home.kpmg/xx/en/home/insights/2019/04/global-construction-survey-future-ready.html

Hahn, E., & López Martín, M. (2015). Robust project management with the tilted beta distribution. *SORT*, *39*, 253-272.

Johnson, D. (1997). The triangular distribution as a proxy for the beta distribution in risk analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *46*(3), 387-398.

Kohonen, T. (2001). *Self-Organizing Maps* (3.ª ed.). Springer-Verlag.

Odeck, J. (2004). Cost overruns in road construction—What are their sizes and determinants? *Transport Policy*, *11*(1), 43-53.

Project Management Institute. (2013). *A guide to the Project Management Body of Knowledge (PMBOK guide) (5th ed.)*. Project Management Institute.

Romero López, C. (2010). *Técnicas de programación y control de proyectos*. Ediciones Pirámide.

Sanchez, F., Bonjour, E., Micaelli, J.-P., & Monticolo, D. (2020). An Approach Based on Bayesian Network for Improving Project Management Maturity: An Application to Reduce Cost Overrun Risks in Engineering Projects. *Computers in Industry*, *119*, 103227.

Schwarz, J. A., & Sánchez, P. M. (2015). Implementation of artificial intelligence into risk management decision-making processes in construction projects. *Universität der Bundeswehr München, Institut für Baubetrieb*, 357-378.

Trietsch, D., Mazmanyan, L., Gevorgyan, L., & Baker, K. R. (2012). Modeling activity times by the Parkinson distribution with a lognormal core: Theory and validation. *European Journal of Operational Research*, *216*(2), 386-396.

Wehrens, R., & Buydens, L. (2007). Self- and Super-organizing Maps in R: The kohonen Package. *Journal of Statistical Software*, *21*, 1-19.

Comunicación alineada con los Objetivos de Desarrollo Sostenible