¿Atraerá el BIG-DATA al P-HACKING al mundo de la ingeniería industrial?

Will BIG-DATA attract P-HACKING to the world of industrial engineering?

Aitor Goti-Elordi[1], Pedro-Ricardo Duarte-Bom[1], José-Antonio Campos-Granados[1] y Diego Galar-Pascual[2, 3]

[1] Universidad de Deusto (España)
[2] Lulea University of Technology (Suecia)
[3] Tecnalia Research & Innovation (España)

DOI: http://dx.doi.org/10.6036/8851

El momento actual en la Industria destaca por la irrupción de términos como Industria 4.0, Sistemas Cyberfísicos, Internet de las Cosas o Big Data. La denominada Industria 4.0 promueve la automatización mediante sistemas informáticos de manufactura para la mejora de sus resultados entre otros con un mejor uso de los datos, pues ésta mejorable gestión de datos ha sido una debilidad habitual de la mejora continua [1]. El Big Data puede ser definido de manera simple como un sistema que permite la colección e interpretación de conjuntos de datos que por su gran volumen no es posible procesarlos con las herramientas convencionales de captura, almacenamiento, gestión y análisis [2]. Los sistemas de Big Data pueden servir, entre otros, para recoger datos, estimar y aplicar más correctamente parámetros de proceso, operación y mantenimiento (como los mostrados p.ej. en [3] [4]).

En este contexto, el disponer de entornos Big Data en la industria abre una vía de acceso a una serie de peligrosas prácticas provenientes de otros sectores. Históricamente, las ciencias no experimentales como la economía, ciertas ciencias sociales, etc. han sido proclives a este tipo de prácticas [5] de gestión al menos dudosa de los datos, principalmente por haber dispuesto de una cantidad de datos suficiente para realizarlas. El presente artículo tiene como objetivo visibilizar esta serie de prácticas estadísticas discutibles para poder velar por una correcta gestión de los estudios basados en entornos Big-Data Industriales.

Los técnicos, científicos e investigadores tienden a publicar o aceptar resultados de estudios alineados con las creencias comunes y a descartar aquellos que ofrecen conclusiones divergentes respecto a las mismas [5]. Como resultado de ello, lo publicado o utilizado no es siempre o totalmente representativo de lo que en la realidad sucede [6][7], llamándosele a este

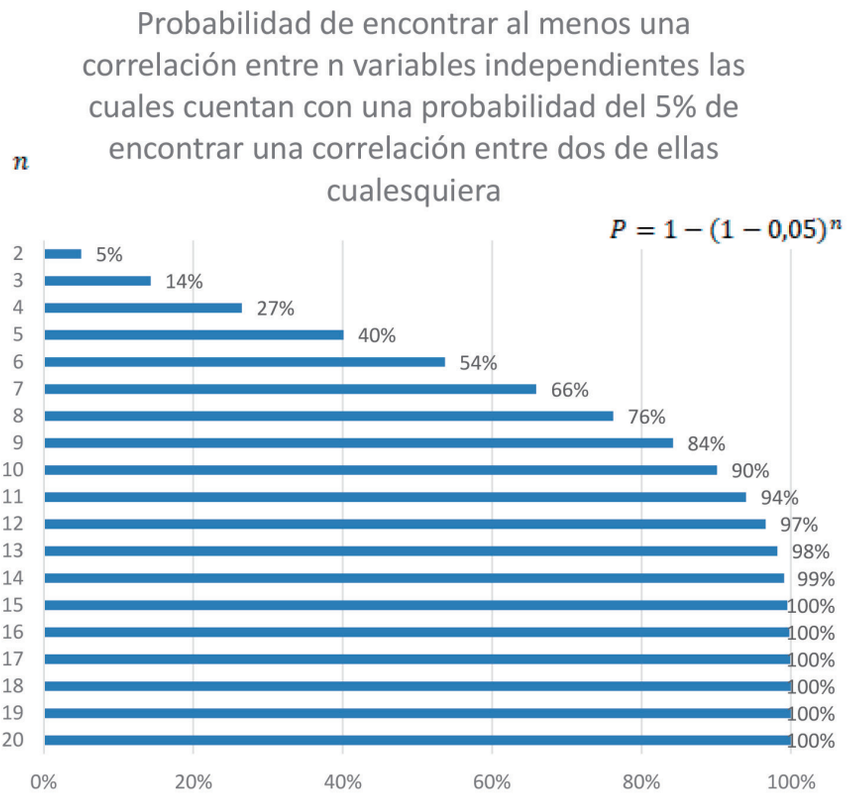

Figura 1: Probabilidad de encontrar al menos una correlación entre n variables independientes las cuales cuentan con una probabilidad del 5% de encontrar una correlación entre dos de ellas cualesquiera

fenómeno sesgo de publicación o Publication Bias. Esto es especialmente problemático cuando los investigadores indagan sobre efectos inexistentes en la realidad, ya que tanto las revistas como los propios autores, mediante el autocensurado o *self-censoring* [8], tienden a que se publiquen a veces las conclusiones erróneas extraídas de datos no representativos de la población y que, por tanto apoyan falsamente su existencia.

La publicación de falsos positivos es destructiva, porque ello sofoca y en ocasiones puede hasta revertir el progreso científico, al menos, hasta que se descubra que las relaciones constatadas de esta manera son falsas. Dentro de este tipo de sucesos convienen destacar dos tipos de prácticas: a) la inicialmente comentada preferencia generalizada por efectos significativos (*selective reporting*) que es compartida (consciente o inconscientemente) entre todos los intervinientes en el proceso científico (autores, evaluadores y editores); y b) la manipulación de datos y modelos en la búsqueda de efectos significativos. En ambos casos la práctica afecta a los investigadores, pero su actuación también afecta a la credibilidad de las publicaciones que recogen este tipo de estudios (y a los pares que las validan), así como al buen fin de las políticas científicas que los financian.

En esta línea una intuición muy popular es que debemos confiar en un resultado si es apoyado por muchos estudios diferentes. Esta intuición se basa en la premisa de que los hallazgos falsos positivos tienden a requerir muchos intentos fallidos cuando, por ejemplo, efectos como a) el 'efecto eminencia' que establece que si un experto lo afirma el dogma no es cuestionado, o b) las presiones de ciertas entidades interesadas en demostrar cosas no totalmente ciertas.

La lógica de confiar en el resultado apoyado por varios estudios diferentes nos lleva a conclusiones erróneas si los científicos explotan la ambigüedad para obtener resultados estadísticamente significativos. Dicho de otra manera, en un entorno en el que se dispone de infinidad de variables y atributos que representan el estado de un sistema en un momento determinado, el investigador puede estudiar correlaciones entre todas ellas hasta encontrar los ele-

mentos que necesita para construir la teoría que desea.

Por ejemplo, como indicaba Gerstman [9] (y se muestra en la Figura 1) si para un caso dado se presentan varias variables independientes con una probabilidad del 5% de encontrar una correlación entre dos de ellas cualesquiera, la probabilidad de encontrar al menos una correlación entre tres variables es del 14%, y con cuatro, cinco, seis, siete, ocho, nueve o diez variables la probabilidad se dispara respectivamente a valores del 27%, 40%, 54%, 66%, 76%, 84% y 90%. El problema se origina cuando en diversas ocasiones estas correlaciones son interpretadas como causalidades, cuando no necesariamente debe existir una relación causa-efecto real.

Al recopilar y analizar datos, los técnicos, científicos y/o investigadores deben tomar infinidad de decisiones como, por ejemplo: recolectar más o menos datos, determinar criterios para incluir o excluir valores atípicos, establecer qué elementos y qué parámetros analizar en ellos, etc. Si estas decisiones no se toman en la fase de diseño de la investigación, sino que se deciden a medida que se analizan los datos, los investigadores pueden sesgar el trabajo (intencionada o no intencionadamente, por ejemplo, para tener más posibilidades de publicarlos científicamente). Este comportamiento es el conocido como *p-hacking* [10] (o *g-bias, significance chasing, data snooping, fiddling, publication bias in situ y specification searching*). Así, *el p-hacking* es la manipulación consciente o inconsciente de datos buscados y obtenidos para una hipótesis con el objetivo de poder o no demostrarla con una determinada confianza y error estadísticos.

El proceso intencionado de *p-hacking* (esto es, el de tratar de mostrar algo que no es cierto como verdadero) implica probar automáticamente un gran número de hipótesis sobre un solo conjunto de datos mediante una búsqueda exhaustiva para encontrar correlaciones (no causalidades). Se considera incluso que la disciplina del *p-hacking* implica buscar distintas pero similares hipótesis, con diferentes variables de control, diferentes instrumentos, diferentes métodos de estimación, diferentes definiciones de variables, etc. Las pruebas convencionales de significación estadística se basan en la probabilidad de que surja un resultado particular si la causalidad solo estuviera en el trabajo, y necesariamente aceptan algún riesgo de conclusiones erróneas de un cierto tipo (rechazos erróneos de la hipótesis nula). Este nivel de riesgo se llama *significancia*. Cuando se realizan grandes cantidades de pruebas, algunas producen resultados falsos de este tipo, por lo tanto, el 5% de las hipótesis elegidas al azar resultan significativas al nivel del 5%, el 1% resultan significativas al nivel de significancia del 1%, y así sucesivamente por la aleatoriedad de los datos. Cuando se prueban suficientes hipótesis, es prácticamente seguro que algunas serán estadísticamente significativas pero engañosas, ya que casi todos los conjuntos de datos con cualquier grado de aleatoriedad probablemente contengan (por ejemplo) algunas correlaciones que no puedan catalogarse como causalidades. Si no se tiene cuidado (o se tiene mala fe), los investigadores que usan técnicas de minería de datos pueden engañar con estos resultados (o ser fácilmente engañados por estos resultados).

A la hora de estudiar el *p-hacking* y cómo puede afectar el mismo a nuestros experimentos e investigaciones, resulta importante distinguir entre datos observacionales y datos experimentales. En áreas en que la investigación se hace mayoritariamente con datos experimentales, hay poco margen para manipulación y ésta se refleja en el diseño del experimento, lo que permite la detección por los revisores o controladores del mismo (salvo que haya manipulación directa de datos). Pero en áreas en que los datos son mayoritariamente observacionales (como, por ejemplo, la economía, determinadas ramas de las ciencias sociales), los grados de libertad para *p-hacking* son casi infinitos, ya que los modelos estadísticos son necesariamente más complejos y tiene que controlar por un gran número de efectos que confunden, "*confounding effects*". Así, es posible encontrar trabajos de estos campos en los que se ponen de manifiesto descubrimientos contradictorios [11], [12].

A priori, puede parecer más complicado que en ingeniería se traten de demostrar hipótesis falsas con significancia alta de una manera creíble. Sin embargo, la Industria 4.0 facilita la disposición de datos observacionales en grandes cantidades, facilitando la entrada al p-hacking en la ingeniería. Los siguientes dos ejemplos simplificados permiten mostrar lo relativamente sencillo que puede llegar a ser aplicar el p-hacking en la industria:

- Si se estudian las producciones de una fábrica que funciona a 3 relevos de 8 horas al día durante 5 días a la semana, pero con datos de una única semana, el analista puede correlacionar y atribuir los malos resultados de un relevo (p.ej. el que se ejecuta entre las 2:00 y 10:00 de la tarde) en cuanto a determinados parámetros (calidad o productividad), cuando tal la relación causa efecto se debería haber atribuido a las condiciones climáticas (p.ej. temperatura) del relevo en cuestión.
- Asimismo, en otro caso similar, en multitud de ocasiones puede resultar fácil correlacionar un aumento del producto defectuoso fabricado con la incorporación de un componente sustitutivo respecto a otro que ha venido utilizándose sin que éste sea la causa raíz que genere dicha consecuencia. La causa raíz del aumento del defectivo puede haber sido otra que haya coincidido en la misma época (p.ej. un aumento de la temporalidad por a) tener que sustituir a trabajadores que se van de vacaciones, o b) un aumento de la saturación de la fábrica).

Por tanto, en este nuevo contexto, ¿qué aportación, además de la de la ética profesional, puede realizar la Ingeniería Industrial para tratar de minimizar la aparición de casos de *p-hacking* en la Industria?

A este respecto merece destacar la conveniencia de analizar cómo ha sido diseñado del estudio. El *p-hacking* resulta del uso de la minería de datos para descubrir patrones en los datos que se pueden presentar como estadísticamente significativos sin primero idear una hipótesis específica en cuanto a la causalidad subyacente, o de realizar una cantidad de experimentos hasta llegar con el conjunto deseado de datos para probar una hipótesis estadísticamente nula ex-ante. Siendo la Ingeniería Industrial capaz de aplicar conocimientos de muy distintas disciplinas, se opina que el ingeniero no debe centrarse únicamente en los datos utilizados para extraer las conclusiones, sino que a) debe fijarse en por qué han sido seleccionados estos datos y no otros, y b) debe exigir cuando sea posible el conjunto de datos completo utilizado en el estudio.

Por tanto, la primera de las recomendaciones, la de cuestionarse si las variables analizadas son las idóneas para el estudio puede llevar al individuo a cuestionarse la lógica del mismo, mientras que la segunda premisa, siempre que sea posible, servirá para realizar contrastes adicionales.

En adición a estas dos premisas principales, en tercer lugar, se recomienda enfrentar la significación estadística con la "significación práctica", en la que se compare no solo si una variable influye en un resultado, sino que se compare su influencia proporcional en comparación con otras.

La investigación científica exige el cumplimiento de dos premisas. La primera es aplicar adecuadamente el método científico, definiendo correcta y exhaustivamente las variables que desean ser analizadas y el tipo de relaciones que se busca en las variables, describiendo las hipótesis a ser testeadas de manera unívoca. En el proceso puede aparecer la "*serendipia*" [13] y descubrirse "*por casualidad*" relaciones no buscadas entre variables, pero si así sucediera, el método científico debiera ser aplicado con rigor a la validación de esta "*sorpresiva*" nueva relación entre variables.

En segundo lugar, es fundamental un comportamiento ético por parte de los técnicos, científicos e investigadores y de los pares que analizan sus descubrimientos. Los investigadores pueden cometer errores, errores que deben ser detectados por los expertos que validan las publicaciones. Pero la publicación de resultados no suficientemente contrastados de una manera intencionada debe ser minimizada tanto desde la autoría como desde la revisión de los mismos. A este respecto, y sobre todo con los científicos de prestigio contrastada, existen barreras reputacionales que reducen dicho riesgo (aunque no lo eliminan, ver el caso de Diederik Stapel [14], por ejemplo). Sin embargo, la gran cantidad de revistas científicas y autores de no tan reconocido prestigio son indicadores que evidencian la necesidad de cuestionarse a veces la validez de un estudio.

En resumen, se puede afirmar que la influencia subjetiva del analista en el resultado del fenómeno a analizar es un elemento a minimizar. El riesgo de que dicha influencia esté presente en los estudios aumenta en los entornos en los que los datos abunden. Hasta hace bien poco esta abundancia de datos se ha encontrado más presente en entornos no ingenieriles, pero en la actualidad y gracias a la Industria 4.0 la realidad es otra. Por tanto, respecto a los dos grandes dominios de clasificación del conocimiento, la tecnología científica y las ciencias sociales (*Science Technology y Social Sciences*, categorizados así por el *Journal Citation Reports*, por ejemplo), el riesgo del *p-hacking* es un elemento a controlar. El presente artículo visibiliza dicho riesgo y emite una serie de acciones para su control. Finalmente, merece destacar que, en los campos tales como la publicación científica se cuenta con las herramientas de meta-análisis que se están desarrollando en la actualidad, las cuales permiten corregir la *Publication Bias*. Así, es de esperar que, en un futuro no muy lejano sea posible aplicar dichas herramientas a cualquier estudio ingenieril.

REFERENCIAS

[1] J. A. Eguren, A. Goti, and L. Pozueta, "Diseño, aplicación y evaluación de un modelo para la mejora de procesos en sectores industriales maduros. Estudio del caso," DYNA Ing. E Ind., vol. 86, no. 3, pp. 59–73, 2011.

[2] A. Goti, A. De la Calle, M. J. Gil, A. Errasti, and J. Uradnicek, "Aplicación de un sistema Business Intelligence en un contexto Big Data de una empresa industrial alimentaria," Dyna, vol. 92, no. 1, pp. 347–353, 2017. DOI: http://dx.doi.org/10.6036/8008

[3] A. Sanchez and A. Goti, "Preventive maintenance optimization under cost and profit criteria for manufacturing equipment, in "Proceedings of ESREL 2006"," 2006, pp. 607–612.

[4] A. Goti, A. Oyarbide-Zubillaga, and A. Sanchez, "Optimizing preventive maintenance by combining discrete event simulation and genetic algorithms," Hydrocarb. Process., vol. 86, no. 10, pp. 115–122, 2007.

[5] R. Rosenthal, "The "File Drawer Problem" and Tolerance for Null Results," Psychol. Bull., vol. 86, no. 3, pp. 638–641, 1979.

[6] J. P. A. Ioannidis, "Why Most Discovered True Associations Are Inflated," Epidemiology, vol. 19, no. 5, pp. 640–648, Sep. 2008.

[7] H. Pashler and C. R. Harris, "Is the Replicability Crisis Overblown? Three Arguments Examined," Perspect. Psychol. Sci., vol. 7, no. 6, pp. 531–536, Nov. 2012.

[8] A. Franco, N. Malhotra, and G. Simonovits, "Social science. Publication bias in the social sciences: unlocking the file drawer.," Science, vol. 345, no. 6203, pp. 1502–5, Sep. 2014.

[9] B. B. Gerstman, Basic Biostatistics. Jones & Bartlett Learning, LLC, 2014.

[10] U. Simonsohn, L. D. Nelson, and J. P. Simmons, "P-curve: A key to the file-drawer.," J. Exp. Psychol. Gen., vol. 143, no. 2, pp. 534–547, 2014.

[11] L. C. MAYES, R. I. HORWITZ, and A. R. FHNSTEIN, "A Collection of 56 Topics with Contradictory Results in Case-Control Research," Int. J. Epidemiol., vol. 17, no. 3, pp. 680–685, Sep. 1988.

[12] T. Kastner, A. Schaffartzik, N. Eisenmenger, K.-H. Erb, H. Haberl, and F. Krausmann, "Cropland area embodied in international trade: Contradictory results from different approaches," Ecol. Econ., vol. 104, pp. 140–144, Aug. 2014.

[13] R. M. Roberts, Serendipia: descubrimientos accidentales de la ciencia RM Roberts - 1991 -. Alianza Editorial, 1991.

[14] Y. Bhattacharjee, "Diederik Stapel's Audacious Academic Fraud -The Mind of a Con Man," The New York Times Magazine, New York , NY, 2013.

AGENDA

PRÓXIMOS EJEMPLARES ESPECIALES

Julio 2019

Tecnología en Edificación

Realizado en colaboración con la Universidad Politécnica de Madrid.

Se tratarán temáticas vinculadas a nuevos materiales, técnicas constructivas, monitorización de indicadores, conductividad térmica, arquitectura bioclimática, códigos técnicos de edificación, entre otras.

Noviembre 2019

Energías alternativas y cambio climático

Realizado en colaboración con la Universidad del País Vasco UPV/EHU y la Universidad de Cantabria.

Se investigarán temas relacionados con energías solares fotovoltaicas y de concentración, energías marinas, energías eólicas, biomasa, hidrogeno como vector energético, almacenaje de energía, vehículos sostenibles, reducciones de emisiones, pilas de combustible, fuentes no contaminantes, etc.