Aplicabilidad de los Métodos de Síntesis Cuantitativa de Experimentos en Ingeniería de Software

E. Fernández[1,2,3], María F. Pollo[2,4], H. Amatriain[2,4], O. Dieste[1], P. Pesado[2,3], R. García-Martínez[2,4]

1 Grupo de Ingeniería de Software Experimental. Facultad de Informática. UPM
2 Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. UNLP
3 Instituto de Investigaciones en Informática LIDI. Facultad de Informática. UNLP - CIC
4 Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. UBA
odieste@fi.upm.es, qiqe2000@hotmail.com, rgarciamar@fi.uba.ar

Resumen. La síntesis cuantitativa [1] consiste en combinar los resultados de varios estudios experimentales con el objeto de generar nuevas piezas de conocimiento. Estas nuevas piezas de conocimientos serán más generales y fiables que los resultados obtenidos por los estudios individuales, ya que dichas piezas de conocimiento están sustentadas por una mayor cantidad de evidencia empírica. El objetivo del presente trabajo es determinar cuáles de los métodos de agregación conviene aplicar en el contexto experimental que hoy día presenta la Ingeniería de Software Experimental.

Palbaras clave. Agregación de experimentos. Síntesis Cuantitativa. Metodo de Monte Carlo. Ingeniería de software Experimental.

1. Introducción

La síntesis cuantitativa [1] consiste en combinar los resultados de varios estudios experimentales con el objeto de generar nuevas piezas de conocimiento. Estas nuevas piezas de conocimientos serán más generales y fiables que los resultados obtenidos por los estudios individuales, ya que dichas piezas de conocimiento están sustentadas por una mayor cantidad de evidencia empírica. En Ingeniería del Software (SE), la síntesis cuantitativa, se ha popularizado en los últimos años desde que fue propuesta por Basilli [2] en 1996 y el primer trabajo de meta-Análisis que se conoce es el desarrollado por Miller [3] que logró combinar 4 experimentos en 1999. A semejanza de lo que se hace en medicina, en SE, se utilizan las Diferencia Medias Ponderadas (WMD) [4] como método de síntesis. Es de hacer notar que para que el método WMD pueda aplicarse de forma fiable requiere que el conjunto de estudios a agregar cumplan ciertas restricciones, entre otras: contener un número mínimo de experimentos, homogéneos y que reportan todos los parámetros estadísticos necesarios (medias, varianzas o desvíos estándar y cantidad de sujetos experimentales). Estas restricciones limitan fuertemente su aplicabilidad en el actual contexto experimental de la SE, donde las replicaciones de experimentos son muy reducidas [5;6] y los estudios no proporcionan los parámetros estadísticos necesarios

debido a problemas de reporte [7;8]. Este hecho provoca disfunciones en la síntesis. Por ejemplo, en [9; 10] si bien se identificó un conjunto de experimentos importante, no se llego a agregar los resultados de los estudios identificados debido a la falta de estandarización de variables respuesta y la baja calidad de los reportes publicados; en [3] solo se pudieron agregar de cuatro estudios experimentales y solo en [11] se logró hacer una síntesis realmente valiosa combinando 15 estudios experimentales.

Si bien WMD es el método de síntesis cuantitativa más difundido y recomendado en la mayoría de las ciencias [12], no es el único que existe. Restringiéndonos únicamente a los métodos cuantitativos, en [4] se propone el método Conteo de Votos Estadístico (SVC) como una alternativa menos restrictiva al WMD, mientras que en [13] se propone el Response Ratio (RR) como método alternativo al WMD. Pero, a la fecha, no se conoce ningún caso de aplicación de estos métodos en SE.

Dado que en algunos trabajos [14; 15] se ha observado que el desempeño de los métodos de Meta-Análisis varía con la cantidad de experimentos incluidos, así como también con la cantidad de sujetos que estos experimentos incluyen, y que en algunos casos el tamaño de la varianza puede cambiar la fiabilidad de los métodos, el objetivo del presente trabajo es determinar cuál o cuáles de los métodos conviene aplicar en el contexto experimental que hoy día presenta la SE. Este trabajo se encuentra estructurado de la siguiente forma: la sección 2 describe en que consiste la síntesis cuantitativa de experimentos y se detallan las principales características de los métodos a evaluar; la sección 3 describe la metodología de análisis a aplicar; la sección 4 se presenta los resultados obtenidos en el proceso de simulación; en la sección 5 detallan las conclusiones obtenidas.

2. Estado de La cuestión

2.1. Síntesis Cuantitativa de estudios Experimentales

La síntesis cuantitativa de estudios experimentales, también llamada agregación de experimentos o Meta-Análisis, consiste en la integración de los resultados de un conjunto de experimentos, previamente identificados, que analizan el desempeño de un par de tratamientos predefinidos con el fin de dar una estimación cuantitativa sintética de todos los estudios disponibles [16]. Como el objeto de estudio de este tipo de trabajo son los estudios experimentales previamente desarrollados y analizados por sus autores, a este tipo de estudio también se lo conoce con el nombre de "Meta-Análisis", que significa "después del análisis" [17].

Si todos los estudios incluidos en el proceso de Meta-Análisis fueran igualmente precisos y utilizaran exactamente las mismas variables respuesta, bastaría con promediar los resultados de cada uno de ellos para obtener así una conclusión final [18]. Sin embargo, en la práctica no todos los estudios tienen la misma precisión, por ello cuando se los combine se debe asignar un mayor peso a los estudios que permiten obtener información más fiable. Esto se logra combinando los resultados mediante un promedio ponderado [19]. Por otra parte, para poder solucionar los problemas vinculados a la no uniformidad de las variables respuesta, los métodos Meta-Analíticos expresan sus resultados mediante un índice de "Tamaño de Efecto", el cual

es un estimador no escalar de la relación entre una exposición y un efecto [16] y es aplicable a cualquier medida de diferencia de los resultados de dos grupos.

El método de síntesis cuantitativa para variables continuas (las mas utilizadas en SE) más utilizado es diferencias medias ponderadas (WMD) [4] (recomendado por organismos internacionales como Cochrane Collaboration [16]). No obstante existen otros métodos alternativos menos difundidos para el cálculo del tamaño de efecto, como son: el Response Ratio (RR) versión paramétrica propuesto por [13], el Vote Counting (VC) propuesto por [4] y el Response Ratio (RR) versión no paramétrica propuesto por [13]. Estos métodos se describen a continuación.

2.1.1 Diferencia Medias Ponderadas

Este método es conceptualmente sencillo: el estimador de efecto individual (representa la tasa de mejora de un tratamiento respecto del otro en cada experimentos) se estima como el cociente de las diferencias entre las medias y el desvío estándar conjunto. Esta función, desarrollada por Glass [17], fue optimizada por Hedgges y Olkin [4] quienes incorporaron un factor de corrección que aumenta la fiabilidad cuando se trabaja con pocos estudios. Convirtiendo a la nueva función en el método de Meta-Análisis mas difundido en la actualidad y el recomendado para ser utilizado en SE [2].

Una vez estimado el tamaño de efecto para cada estudios, puede estimarse el *efecto global* (representa la tasa de mejora de un tratamiento respecto del otro a nivel general) el cual se calcula como una media ponderada de los estimadores de efecto de los estudios individuales [4]:

$$d^* = \frac{\sum d_i / \sigma^2{}_i(d)}{\sum 1/\sigma^2{}_i(d)} \qquad \textbf{(1)}$$

d^* es el tamaño de efecto global
$\sum d_i / \sigma^2{}_i(d)$ es la suma de los efectos individuales
$\sum 1/\sigma^2{}_i(d)$ es la suma de la inversa varianza

Para mayores detalles de cómo aplicar las formulas indicadas remitirse a [4].

2.1.2 Response Ratio Paramétrico (PPR)

El PRR consiste en estimar un índice de efecto, o Ratio, entre dos tratamientos mediante el cociente de ambas medias [20]. Este cociente estima la proporción de mejora existente entre ambos tratamientos [21]. Así, por ejemplo, un ratio de 1.3 indicará que el tratamiento principal es un 30% mejor que el secundario, o un ratio de 1 indicará que no hay diferencias en el desempeño de ambos tratamientos.

La aplicación del método es similar a WMD. Primeramente se debe estimar el Ratio de cada uno de los experimentos ($RR = Y^E / Y^C$) y luego, en base a estos, se estima el Ratio global mediante un promedio ponderado de los ratios individuales:

$$L^* = \frac{\sum_{i=1}^{k} W_i^* L_i}{\sum_{i=1}^{k} W_i^*} \qquad \textbf{(2)}$$

L^* es el efecto global
L_i es el efecto de cada estudio
W_i es el factor de peso = $1/v$

Donde cada estudio es ponderado en base a la inversa de su varianza:

$$v = \frac{S^{2E}}{n^E Y^E} + \frac{S^{2C}}{n^C Y^C}$$

v es el error típico
S^2‘s son las varianzas de los estudios
Y‘s son las medias de los estudios
n‘s son las cantidades de sujetos

(3)

Para que la combinación de un conjunto de estudios sea más precisa se incorporó al método el logaritmo natural, el cual aplicado a los efectos de los estudios individuales permite linealizar los resultados y normalizar su distribución. Para mayores detalles de cómo aplicar las formulas indicadas remitirse a [21].

2.1.3 Conteo de Votos Estadístico (SVC)

El SVC es un método que requiere muy poca información para poder ser aplicado. Solo precisa conocer si existe o no diferencia entre las medias de los tratamientos (a lo cual llamaremos “voto”) y la cantidad de sujetos experimentales utilizados en cada estudio (utilizado como ponderador del “voto”) [4]. En base a estos datos se realiza un proceso de inferencia estadística con el objeto de determinar que tamaño de efecto (en general seleccionado de una lista que va desde -0,5 a 0,5) tiene la mayor probabilidad de ser el tamaño de efecto real que se hubiera estimado mediante WMD si se contara con todos los datos para poder hacerlo. La función principal de estimación es:

$$L(\delta \mid X_1,.....,X_i) = \sum_{i=1}^{k} \begin{cases} X_i \ln\left[1-\phi\left(-\sqrt{\tilde{n}}\delta\right)\right] + \\ (1-X_i)\ln\phi\left(-\sqrt{\tilde{n}}\delta\right) \end{cases}$$

$L(\delta|X_1,....X_n)$ es la probabilidad de tamaño de efecto
δ es el tamaño de efecto a testear
X_i es el valor del voto de cada estudio
ñ = $(n^E+n^C) / (n^E*n^C)$ donde n’s son las cantidades de sujetos experimentales de cada estudio

(4)

Para mayores detalles de cómo aplicar las formulas indicadas remitirse a [4].

2.1.4 Response Ratio No Paramétrico

Esta versión del RR es similar a la versión paramétrica, siendo su principal diferencia la forma en que pondera a los estudios. En lugar de utilizar la inversa de la varianza, el NPRR utiliza la cantidad de sujetos experimentales [21]:

$$v = \frac{n_C + n_E}{n_E n_C} + \frac{Ln(RR^2)}{2(n_C + n_E)}$$

v es el error típico
n‘s son las cantidades de sujetos
RR es el Ratio

(5)

La principal ventaja de este método, desde el punto de vista de su aplicación, consiste en no requerir conocer las varianzas de los tratamientos ni requerir que exista normalidad y homeosticidad, lo cual aumenta enormemente sus probabilidades de aplicación. Para mayores detalles de cómo aplicar las formulas indicadas remitirse a [21].

3. Metodología

El objetivo de este trabajo es desarrollar un proceso de simulación, complementario a los anteriores [14, 15], para evaluar el desenvolvimiento de los

cuatro métodos de Meta-Análisis (WMD, PRR, SVC y NPRR) de forma exhaustiva (variando: los tamaños de efectos y la cantidad de experimentos, la cantidad de sujetos por experimento), en un contexto experimental como el que hoy día presenta la SE. Donde, si bien la cantidad de experimentos ha venido creciendo significativamente en los últimos años (pasando de aproximadamente tres estudios experimentales por año a principio de los 90' a más de veinte estudios por año, publicados en los principales congresos y revistas, a principio del años 2.000 [22]), todavía existen pocos experimentos y los mismos, en general, utilizan pocos sujetos experimentales (por ejemplo: [23] utiliza 4 sujetos experimentales) y tienen falencias en sus reportes (es común que no se publiquen las varianzas de los tratamientos analizados como sucede por ejemplo en: [24]).

De forma similar a como se hizo en [14] y [15], para desarrollar el proceso de simulación utilizaremos la técnica de Monte Carlo. La simulación de Monte Carlo es una técnica que combina conceptos estadísticos (muestreo aleatorio) con la capacidad que tienen los ordenadores para generar números pseudo-aleatorios siguiendo una distribución de probabilidad normal. En este contexto, se utilizó esta técnica para simular los valores que hubieran generado los distintos sujetos en la aplicación de los tratamientos, en base a los cuales, luego, se estimaraon la media y la varianza de cada experimento.

El primer paso del desarrollo del proceso de simulación consiste en definir los valores poblacionales a partir de los cuales se obtienen los valores de la muestra para simular el Meta-Análisis. Para ello se siguieron las recomendaciones de los anteriores trabajos de simulación [14] y [15] e información relevante de cómo son típicamente los estudios hechos en el ámbito de la. Los tamaño de efecto poblacional a analizar son los mismo que se definen en [21] bajo (0,2), medio (0,5) y alto (0,8), mas la incorporación del tamaño de efectos muy alto (1,2), ya que se ha observado que varios estudios [25; 26] hechos en SE tienden a dar tamaños de efectos muy altos.

La media poblacional del tratamiento secundario (μ^c) es fijada en 100 y los desvíos estándar como se hizo en [21] son fijado en los siguientes porcentajes respecto de la media de dicho tratamiento: 10% al cual llamaremos varianza baja; 40% al cual llamaremos varianza media; y 70% al cual llamaremos varianza alta. Por su parte la media poblacional del tratamiento principal se estimará de la siguiente forma $\mu^E = 100 + \delta * \sigma$ y el ratio poblacional que se utilizará para validar los resultados que generen el RR paramétrico y no paramétrico será estimado: RR = μ^E / μ^c.

Por otra parte, la cantidad de experimentos a agregar en cada proceso de agregación irá desde 2 a 10 incrementándose de dos en dos, por considerar que el contexto experimental de la SE no aporta hoy día muchos experimentos potencialmente agregables mediante Meta-Análisis, tal y como puede comprobarse en las revisiones sistemáticas hechas hasta el momento [9; 10]. De igual modo, consideramos un número reducido se sujetos por experimento, en el rango de 4 – 20.

Las variables respuesta resultado de la simulación serán también las utilizadas en los estudios de Lajenouse [14] y Friedrich y colegas [15], esto es, el error de tipo I y II ó, para ser mas exacto, su inversa: (1-α) o precisión y (1 - β) o poder estadístico. Estas variables son adecuadas para nuestro propósito ya que determinan cuantas veces

se equivoca un método de Meta-Análisis a la hora de determinar si existe (error de tipo I) o no (error de tipo II) una diferencia significativa entre dos tratamientos. Por último, siguiendo las recomendaciones de [14] para cada combinación de valores de las variables se construirán 1.000 simulaciones, tras lo cual se calcularan los valores de las variables respuesta.

4. Resultados

Los resultados detallados de la simulación se presentan en las tablas 1 a 6 (al final del trabajo). Las tablas vinculadas a la fiabilidad indican el porcentaje de veces que el intervalo de confianza estimado (a un nivel de $\alpha = 0.05$) contuvo el valor del tamaño de efecto poblacional, mientras que las tablas vinculadas a la potencia estadística indican el porcentaje de veces que dicho intervalo de confianza no contuvo el valor 0 para los métodos WMD y SVC y el valor 1 para los métodos RR paramétricos y no paramétricos. Para facilitar la compresión de las mismas, se han resaltado las celdas en las cuales los porcentajes estimados superaban al valor mínimo fijado, $1 - \alpha = 95\%$ para la fiabilidad y $1 - \beta = 80\%$ (el cual es el valor típicamente recomendado [27]) para la potencia estadística.

Respecto del desempeño de cada uno de los métodos podemos decir que:

- Es fiable utilizar el método WMD en contextos experimentales donde los tamaños de efecto poblacionales son bajos o medios, siendo su condición óptima de aplicación cuando los efectos son medios y el conjunto de experimentos a agregar superen a los 112 sujetos experimentales. Cuando los efectos poblacionales son altos o muy altos, el método tiende a perder fiabilidad sobre todo cuando se incrementa la cantidad de experimentos y la cantidad de sujetos experimentales.
- Es aconsejable utilizar el método PRR, siempre y cuando los estudios a agregar posean más de 4 sujetos experimentales. El método mostró ser robusto ante los cambios en la varianza, tamaños de efecto y cantidad de experimentos a agregar. Su condición óptima de aplicación varía en función del tamaño de efecto poblacional y la cantidad de sujetos experimentales que los estudios totalicen, observando que: para efectos muy altos se requieren por lo menos 80 sujetos experimentales, para efectos altos se requieren como mínimo 100 sujetos experimentales y para un efecto medio se requieren como mínimo 140 sujetos experimentales, para que el método posea fiabilidad y potencia estadística.
- Es fiable utilizar el método SVC, solo cuando el tamaño de efecto es medio se cuenta con experimentos que totalicen más de 80 sujetos experimentales. Su falta de fiabilidad es compensada en parte con su alta potencia, pero se debe tener mucho cuidado con el uso del mismo sobre todo en contextos experimentales donde el tamaño de efecto poblacional es bajo. En contextos de tamaños de efectos altos, la perdida de fiabilidad es compensada en parte con la alta potencia estadística.
- El método NPRR ha sido el método más fiable de todos los analizados. Su mayor problema está dado por la baja potencia estadística que se acentúa en contextos donde la población tiene baja varianza. Esto se debe en parte a que en contexto de baja varianza no se requiere que la diferencia entre las medias sea excesiva para que el efecto sea alto. Su condición óptima de aplicación varía en función de la varianza

poblacional, el tamaño de efecto poblacional y la cantidad de sujetos experimentales que los estudios totalicen observando que: para varianzas poblacionales medias y tamaños de efecto poblacionales altos o muy altos se requieren como mínimo 100 sujetos experimentales, para varianzas poblacionales altas con tamaños de efecto poblacionales muy altos se requieren como mínimo 48 sujetos experimentales, para efectos poblacionales medios se requieren como mínimo 80 sujetos experimentales y para efectos poblacionales altos se requieren como mínimo 16 sujetos experimentales, para que el método posea fiabilidad y potencia estadística.

5 Conclusión

A modo de conclusion preliminar podemos decir que dentro de los parámetros normales que hoy presenta la Ingeniería de Software Empírica [5] el método WMD ha mostrado comportarse de forma confiable, por lo que no es necesario utilizar el método PRR como método alternativo al mismo. Por otra parte, en los casos en que los reportes experimentales no sean completos, el método NPRR mostró un comportamiento mucho mas fiables que el SVC que, en general, no dio buenos resultados.

No obstante esto, si se trabaja en un entorno donde los tamaños de efecto son altos, el contexto cambia drásticamente, ya que aquí el método WMD deja de ser fiable, lo cual implica que los tamaños de efectos estimados pueden no ser correctos, por tal motivo el método PRR, que si ha mostrado ser fiable cuando los tamaños de efecto son altos, se convierte en el método mas recomendable cuando los reportes son completos, mientras que el método NPRR sigue siendo el mejor método cuando los reportes no son completos.

6 Referencias

1. Goodman C.; 1996; Literature Searching and Evidence Interpretation for Assessing Health Care Practices; SBU; Stockholm.
2. Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sörumgård, S., Zelkowitz, M.; 1996; *The empirical investigation of perspective-based reading*, International Journal on Empirical Software Engineering, Vol. 1, No. 2; pp. 133–164.
3. Miller, J; 1999: Can Results from Software Engineering Experiments be Safely Combined? IEEE METRICS, 152-158
4. Hedges, L.; Olkin, I.; 1985; *Statistical methods for meta-analysis*. Academic Press.
5. Tonella P., Torchiano M., Du Bois B., Systä T.; 2007; *Empirical studies in reverse engineering: state of the art and future*
6. Sjoberg, D.; 2005; A survey of controlled Experiments in Software Engineering; ; IEEE Transactions on Software Engineering; Vol 31 Nro. 9
7. Biffl, S.; Halling,M.; Köszegi S.; 2003; *Investigating the Accuracy of Defect Estimation Models for Individuals and Teams Based on Inspection Data*; Proceedings of the 2003 International Symposium on Empirical Software Engineering (ISESE'03)

8. Fusaro, P., Lanubile, F., Visaggio, G.; 1997; A replicated experiment to assess requirements inspection techniques; 1997 Empirical Software Engineering, 2, 39-57 (1997)
9. Juristo N., Moreno A., Vegas S.; 2004; Towards building a solid empirical body of knowledge in testing techniques. ACM SIGSOFT Software Engineering Notes (SIGSOFT) 29(5):1-4
10. Jřrgensen, M.; 2004; A *Review of Studies on Expert Estimation of Software Development Effort*. Journal of Systems and Software. (70): 1-2, pp. 37-60.
11. Dyba, T., Aricholm, E.; Sjoberg, D.; Hannay J.; Shull, F.; 2007; Are two heads better than one? On the effectiveness of pair programming. IEEE Software;12-15.
12. Shercliffe, R.; Stahl, W.; Tuttle, M.; 2009; *The Use of Meta-analysis in Psychology*; Theory & Psychology, Vol. 19, No. 3, 413-430 (2009)
13. Gurevitch, J. and Hedges, L.; 2001; *Meta-analysis: Combining results of independent experiments.* Design and Analysis of Ecological Experiments (eds S.M. Scheiner and J. Gurevitch), pp. 347–369. Oxford University Press, Oxford.
14. Lajeunesse, M & Forbes, M.; 2003; *Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques*. Ecology Letters, 6: 448-454.
15. Friedrich, J, Adhikari, N; Beyene, J; 2008; The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study; BMC Medical Research Methodology
16. Cochrane; 2008; Curso Avanzado de Revisiones Sistemáticas; www.cochrane.es/?q=es/node/198
17. Glass, G; 1976; Primary, secondary, and meta-analysis of research. Educational Researcher 5: 3-8
18. Borenstein, M.; Hedges, L; Rothstein, H.; 2007; Meta-Analysis Fixed Effect vs. random effect; www.Meta-Analysis.com
19. Kitchenham, B. A.; 2004; *Procedures for performing systematic reviews*. Keele University; TR/SE-0401. Keele University Technical Report.
20. Worn, B.; Barbier, E.; Beaumont, N.; Duffy, J.; Folke, C; Halpern, B.; Jackson, J.; Lotze, H.; Micheli, F.; Palumbi, S.; Sala, E.; Selkoe, K.; Stachowics, J.; Watson, R; 2007; Supporting Online Material: Impacts of biodiversity loss on ocean ecosystem services.
21. Miguez, E. & Bollero, G; 2005; Review of Corn Yield Response under winter cover cropping systems using Meta-Analytic Methods; Crop Science Society of America
22. Sjoberg, D.; 2005; A survey of controlled Experiments in Software Engineering; ; IEEE Transactions on Software Engineering; Vol 31 Nro. 9
23. Burton, A., Shadbolt, N., Rugg, G. y Hedgecock, A.; 1990. *The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Level of Expertise*. Knowledge Acquisition 2(2): 167-178.
24. Denger, C; Ciolkowski M; Lanubile, F; *Does Active Guidance Improve Software Inspections? A Preliminary Empirical Study*; 2004; Proceedings of the IASTED International Conference SOFTWARE ENGINEERING February 17-19, 2004, Innsbruk, Austria; 408-413
25. Corbridge, C., Rugg, G., Major, P., Shadbolt, N. y Burton, A. 1994. *Laddering: Technical and Tool in Knowledge Acquisition*. Department of Psychology, University of Nottingham; Nottingham NG7 2RD.
26. Woody, J.; Will, R.; Blanton, J.; *Enhancing Knowledge Elicitation using the Cognitive Interview;* Expert system with application; 1996; Vol. 10 N. 1
27. Cohen, J.; *Statistical Power Analysis for the Behavioral Sciences.* (2nd ed.) 1988. ISBN 0-8058-0283-5.

Tabla 1. Comparación de la fiabilidad de los métodos de agregación

			WMD					RRP					VCE					RRNP				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Fiabilidad -Varianza baja	EF: 0,2	4	98,4	100	100	99	100	93,9	89,7	85,5	93,3	93,5	0,9	0,7	0,3	0,3	0,2	100	100	100	100	100
		8	98,1	100	100	100	100	98,1	100	100	100	100	1,2	0,5	0,2	0,3	0,2	100	100	100	100	100
		10	97,6	100	100	100	100	96,6	99,3	100	100	100	0,7	0,4	0,3	0,3	0	100	100	100	100	100
		14	97,6	100	100	100	100	97,6	97,6	100	100	100	1,2	0,5	0,3	0,3	0,3	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	1,1	0,5	0,3	0,3	0,2	100	100	100	100	100
	EF: 0,5	4	97,3	100	100	99,4	96,3	91,6	90,2	89,5	92,3	93	57	75,6	85,6	92,6	79	100	100	100	100	100
		8	96,3	100	98,9	98,1	98,1	97,5	100	100	100	100	53,7	80,8	83,7	92,7	97,2	100	100	100	100	100
		10	92,1	97,9	100	99	93,1	96,5	100	100	100	100	68,9	86,5	89,7	93,9	94	100	100	100	100	100
		14	97,4	99,4	98,5	98,4	98,2	96,1	98,3	100	100	100	83	92,1	99	100	100	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	92,9	98,9	99,1	98,6	100	100	100	100	100	100
	EF: 0,8	4	95,6	99,1	99,1	97,1	94,1	93,6	90,4	90,8	94,2	95,2	0,6	0,4	0,2	0,2	0,2	100	100	100	100	100
		8	93,7	95,6	96	92,8	93,4	96,3	100	100	100	100	0,5	0,4	0,2	0,2	0,2	100	100	100	100	100
		10	87,6	92,1	95,9	92,2	83,5	96	100	100	100	100	0,6	0,4	0,2	0,2	0,2	100	100	100	100	100
		14	98,4	91,8	94,2	90,5	90,3	96,5	98,3	100	100	100	0,6	0,4	0,2	0,2	0,2	100	100	100	100	100
		20	100	100	96,4	92,5	100	100	100	100	100	100	0,6	0,4	0,2	0,2	0,2	100	100	100	100	100
	EF: 1,2	4	95,4	96,3	92,4	88,1	82,2	91,9	91,6	89,6	94,3	95	0,2	0,2	0	0	0	100	100	100	100	100
		8	90,5	91,4	83,8	73,1	80,7	97,3	100	100	100	100	0,2	0,2	0	0	0	100	100	100	100	100
		10	81,3	81,1	85,7	79,5	62,8	96,1	100	100	100	100	0,2	0,2	0	0	0	100	100	100	100	100
		14	94,7	79,1	81,6	68,3	55	96,3	97,6	100	100	100	0,2	0,2	0	0	0	100	100	100	100	100
		20	98,2	91,6	72,5	63,1	51,5	100	100	100	100	100	0,2	0	0	0	0	100	100	100	100	100

Tabla 2. Comparación de la potencia estadística de los métodos de agregación

			WMD					RRP					VCE					RRNP				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Potencia – Varianza baja	EF: 0,2	4	1,0	2,0	1,1	1,2	2,7	3,9	5,1	10,3	9,1	11,8	40,9	51,6	57,8	65,6	70,0	0,0	0,0	0,0	0,0	0,0
		8	2,8	5,2	1,2	3,1	5,9	2,8	7,7	2,6	4,2	7,7	37,5	59,7	73,9	77,3	87,7	0,0	0,0	0,0	0,0	0,0
		10	6,9	7,1	1,9	8,5	11,9	8,9	7,9	4,9	10,1	16,5	47,3	62,7	64,9	75,7	81,2	0,0	0,0	0,0	0,0	0,0
		14	8,9	7,7	9,7	1,8	13,1	8,9	11,7	15,1	5,9	18,3	38,6	55,5	74,0	77,9	87,9	0,0	0,0	0,0	0,0	0,0
		20	4,6	1,7	0,0	19,1	0,0	4,6	1,7	0,0	22,2	5,9	54,4	77,0	81,5	83,8	100	0,0	0,0	0,0	0,0	0,0
	EF: 0,5	4	5,0	8,5	16,5	19,8	25,0	9,7	24,1	33,6	40,8	47,9	56,0	75,1	85,1	92,6	93,7	0,0	0,0	0,0	0,0	0,0
		8	5,6	25,8	45,2	61,1	82,6	15,6	34,4	56,7	71,2	85,7	53,1	80,3	83,2	92,7	97,2	0,0	0,0	0,0	0,0	0,0
		10	28,2	40,8	52,5	77,9	84,8	32,2	44,3	66,0	84,7	90,8	68,4	86,0	89,3	93,9	94,0	0,0	0,0	0,0	0,0	0,0
		14	31,5	57,6	83,6	98,4	100	36,8	60,1	92,7	98,4	100	82,3	91,6	98,5	100	100	0,0	0,0	0,0	0,0	0,0
		20	41,9	100	100	100	100	47,7	100	100	100	100	92,2	98,4	99,1	98,6	100	0,0	0,0	0,0	0,0	0,0
	EF: 0,8	4	11,6	32,5	44,6	70,0	79,8	22,4	42,0	65,8	80,0	90,3	78,1	94,6	99,0	100	100	0,0	0,0	0,0	0,0	0,0
		8	33,4	70,9	94,4	98,9	100	43,1	78,5	95,2	100	100	78,6	93,9	94,5	100	100	0,0	0,0	0,0	0,0	0,0
		10	53,4	81,2	100	100	100	54,4	88,2	100	100	100	85,0	97,6	99,8	100	100	0,0	0,0	0,0	0,0	0,0
		14	66,3	98,0	100	100	100	67,8	98,0	100	100	100	99,8	99,8	99,8	100	100	0,0	0,0	0,0	0,0	0,0
		20	97,7	100	100	100	100	97,7	100	100	100	100	99,8	99,8	100	100	100	0,0	0,0	0,0	0,0	0,0
	EF: 1,2	4	30,9	73,3	95,1	98,8	98,1	46,3	88,5	97,5	98,8	99,2	94,0	99,8	99,8	100	100	0,0	0,0	0,0	0,0	0,0
		8	73,5	100	100	100	100	77,5	100	100	100	100	94,1	99,2	99,8	100	100	0,0	0,0	0,0	0,0	0,0
		10	78,8	100	100	100	100	81,3	100	100	100	100	99,8	99,8	99,8	100	100	0,0	0,0	0,0	0,0	0,0
		14	100	100	100	100	100	100	100	100	100	100	99,8	99,8	99,8	100	100	0,0	0,0	0,0	0,0	0,0
		20	100	100	100	100	100	100	100	100	100	100	99,8	99,8	100	100	100	0,0	0,0	0,0	0,0	0,0

Tabla 3. Comparación de la fiabilidad de los métodos de agregación

			WMD					RRP					VCE					RRNP				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Fiabilidad -Varianza media	EF: 0,2	4	97,8	100	100	99,2	100	96,4	92	90	93,2	92	0,8	0,6	0,4	0,2	0,4	100	100	100	100	100
		8	97	100	100	100	100	97	100	100	100	100	1	0,8	0,4	0,2	0,2	100	100	100	100	100
		10	97,4	100	100	100	100	96	100	100	100	100	1	0,4	0,4	0,4	0,2	100	100	100	100	100
		14	96	100	100	100	100	96	98,6	100	100	100	1	1	0,2	0,2	0,2	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	1	0,4	0,2	0,2	0,2	100	100	100	100	100
	EF: 0,5 - 0,8	4	97	100	100	99,1	96,6	90,7	92,3	90	93,6	94,3	56,7	76	85,4	91	80,6	100	100	100	100	100
		8	96,3	100	98,6	96,8	97,8	97,2	100	100	100	100	52,9	81	85,2	91,7	97	100	100	100	100	100
		10	93,3	97,7	100	98,9	91,8	95,7	100	100	100	100	68,1	85,9	89,9	94,6	94,4	100	100	100	100	100
		14	97,6	99,5	98,7	99,1	98	96,2	98,8	100	100	100	82,4	90	99	100	100	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	92,5	97,8	98,5	99	100	100	100	100	100	100
	EF: 0,8	4	96,7	98,8	98,5	95,5	95	93,5	90,6	93,3	92,2	90,7	0,2	0,2	0,1	0,1	0,1	100	100	100	100	100
		8	94,7	95,3	94,6	90,7	93,9	97,5	100	98,7	98,7	100	0,3	0,2	0,1	0,1	0,1	100	100	100	100	100
		10	85,7	91,9	95,4	95	81,7	95,6	99,4	100	100	100	0,3	0,2	0,1	0,1	0,1	100	100	100	100	100
		14	96,9	92,2	95,6	89,7	88,9	98,4	100	99,3	100	100	0,3	0,2	0,1	0,1	0,1	100	100	100	100	100
		20	100	100	94,8	89	100	100	100	100	100	100	0,3	0,2	0,1	0,1	0	100	100	100	100	100
	EF: 1,2	4	96,7	94,1	92,6	87,9	81,6	93,7	89,9	92,8	92,7	90,8	0,1	0,1	0	0	0	100	100	100	100	100
		8	92	91,7	81,6	76,8	79,7	97,5	100	98,8	99,2	100	0,1	0,1	0	0	0	100	100	100	100	100
		10	81,5	81,5	84,9	78,8	59,8	96,8	99,4	100	100	98,6	0,1	0,1	0	0	0	100	100	100	100	100
		14	94,8	79,2	80,3	60,9	55,4	99,4	100	99,8	100	100	0,1	0	0	0	0	100	100	100	100	100
		20	97,2	91,9	67,8	58,6	47,9	100	100	100	100	100	0,1	0	0	0	0	100	100	100	100	100

Table 4. Comparación de la potencia estadística de los métodos de agregación

			WMD					RRP					VCE					RRNP				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Potencia – Varianza media	EF: 0.2	4	0.4	1.6	1.8	1.2	1.6	0.6	3.6	5	4.6	8	37.8	51.4	54.4	68.2	68.4	0	0	0	0	0
		8	2.6	5.2	0.6	2.4	5.4	2.6	4.4	0.6	1.8	2.6	37.2	55.6	70.2	77.2	87.6	0	0	0	0	0
		10	7.8	9.2	3	7	11.6	7.8	6.8	1.8	5.6	9.4	49.8	59.8	65.2	78.8	80.6	0	0	0	0	0
		14	8.6	6.8	8.6	1	16.8	8.4	6	6.4	1	10.6	42.4	55.2	76	78.8	88.4	0	0	0	0	0
		20	2.8	2	0	16.4	0	2.8	1.2	0	15.6	0	53.4	76.8	81	84.6	100	0	0	0	0	0
	EF: 0.5	4	5.3	10.3	16.4	17.7	25.3	6.1	18.4	22.3	23.4	34.8	56.1	75.7	85.1	91	93.2	0	0	0	0	0
		8	5.1	25.4	48.4	59.3	81.1	6.8	25	42.4	53	74.8	52.4	80.7	84.9	91.7	97	0	0	0	0	0
		10	27.9	39.5	55	78.3	84.7	28	37.7	50.2	79.8	76.1	67.8	85.6	89.6	94.6	94.4	0	0	0	0	0
		14	32.1	57.1	83.7	99.1	100	31	53.3	82.9	97.3	99	82.1	89.7	98.7	100	100	0	0	0	0	2
		20	44.9	100	100	100	100	42.5	97.6	100	100	100	92.2	97.5	98.5	99	100	0	0	0	0.9	0
	EF: 0.8	4	13	33.2	44.2	68.6	84.1	18.2	35.2	49.5	68.6	76.2	77.6	95.6	99.2	100	100	0	0	0	1.1	1.9
		8	37.1	70.7	93.6	98.3	100	36.1	72.8	91.9	98.7	100	79.1	94.9	94.3	100	100	0	0	0	8.6	19.6
		10	54	81.1	100	100	100	52.8	80.8	100	100	100	83	97.5	99.9	100	100	0	1.3	1.5	28.1	56.4
		14	66.8	98.7	100	100	100	59.2	93.4	100	100	100	99.9	99.9	99.9	100	100	0	4.8	31.8	59.2	94.1
		20	97.6	100	100	100	100	97.6	100	100	100	100	99.4	99.9	100	100	100	0	7.8	81.3	100	100
	EF: 1.2	4	33	75.8	96.6	99.3	97.8	39.7	79.2	97.9	99.3	97.8	94.3	99.9	99.9	100	100	0	0	0.8	5.5	20.3
		8	76.9	100	100	100	100	75.3	100	100	100	100	96.7	99.6	99.9	100	100	0	5.6	32.9	69.7	96.1
		10	79.5	100	100	100	100	80.5	100	100	100	100	99.9	99.9	99.9	100	100	0	20.1	68.7	95.6	100
		14	100	100	100	100	100	100	100	100	100	100	99.9	99.9	99.9	100	100	5	54.2	98.9	100	100
		20	100	100	100	100	100	100	100	100	100	100	99.9	99.9	100	100	100	16.6	100	100	100	100

Table 5. Comparación de la fiabilidad de los métodos de agregación

			WMD					RRP					VCE					RRNP				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Fiabilidad -Varianza alta	EF: 0.2	4	98.2	100	100	98.7	100	96.2	90.2	89.7	89.7	87.2	0.5	0.3	0.1	0.2	0.1	97.5	96.5	99.3	95.8	94.5
		8	95.8	100	100	100	100	99	100	100	100	100	0.6	0.4	0.1	0.1	0.1	97.4	100	98.8	100	100
		10	97.2	100	100	100	100	95.7	100	100	100	100	0.6	0.2	0.1	0.3	0.1	99.2	100	100	100	100
		14	96.9	100	100	100	100	98.5	100	100	100	100	0.7	0.3	0.1	0.1	0.1	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	0.4	0.2	0.1	0.2	0.1	100	100	100	100	100
	EF: 0.5 - RR	4	97.7	100	100	98.5	96.3	94.6	93.2	92.8	88.8	83	59	76	84.1	90.5	77.2	98.8	100	100	97.5	96.3
		8	95.2	100	99.6	95.8	97.5	96.6	99	98.3	95.9	100	54.3	77.8	87.8	89.2	97.2	97.8	100	100	100	100
		10	92.3	98.4	100	98.9	93.8	96.9	99.2	100	98.4	98.1	70.8	85.5	89.1	96	94.3	100	100	100	100	100
		14	97.6	98.8	97.7	98.6	98	100	98	98.9	100	100	80.3	89.7	98.6	100	100	100	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	91.9	96.8	98.8	99	100	100	100	100	100	100
	EF: 0.8	4	96.4	99.1	99.1	96.2	94	95.3	93.4	94.4	89.7	87.6	0.4	0.2	0.1	0.1	0.1	98.8	100	100	99.5	100
		8	94.2	95.7	96.9	92.5	94.2	95.5	94.2	97.2	96.3	98	0.3	0.2	0.1	0.1	0.1	99	100	100	100	100
		10	86.4	91.8	96.7	93.3	79.6	95.1	98.6	98.4	96.4	94.2	0.3	0.2	0.1	0.1	0.1	100	100	100	100	100
		14	97.7	92.5	95.6	91.6	88.9	100	96.3	99.3	100	99	0.3	0.2	0.1	0.1	0.1	100	100	100	100	100
		20	100	100	94.7	90.1	100	100	100	100	100	100	0.3	0.2	0.1	0.1	0.1	100	100	100	100	100
	EF: 1.2	4	94.9	96.4	93.6	89.7	80.4	95	90.6	92.1	89.3	84.7	0.1	0.1	0	0	0	98.8	100	100	99.4	100
		8	90.6	91.4	80.4	73.4	80.3	95.1	95.3	94.7	94	97	0.1	0.1	0	0	0	99	100	100	100	100
		10	81.8	83.2	85.3	81.4	60.1	94.6	95.6	97.2	96.6	87.1	0.1	0.1	0	0	0	100	100	100	100	100
		14	96.1	80	80.1	66.2	57.9	100	95.8	99.4	99	97.1	0.1	0.1	0	0	0	100	100	100	100	100
		20	98.1	91.7	67.2	60.7	52.3	100	100	100	100	100	0.1	0	0	0	0	100	100	100	100	100

Table 6. Comparación de la potencia estadística de los métodos de agregación

			WMD					RRP					VCE					RRNP				
			2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Potencia – Varianza alta	EF: 0.2	4	1.1	1.8	0.7	0.7	1.4	2.2	3.5	4.2	5.1	2.2	41.1	50.2	59.6	64.9	69.5	1.6	0.5	0.5	1.7	1.4
		8	2.2	5.2	0.9	3.1	6	0	1.3	0	1	2.6	40.4	59.8	71.4	75.3	87.3	0.9	0	0	0	2.6
		10	10	10.1	1.9	8.3	13.7	2.7	6.6	0.9	6.5	5.8	52.1	64.4	63.8	74.8	80.9	0	1.2	0	1.1	3.2
		14	10.4	8.3	9.9	2.3	14.1	5.9	3.2	4.2	1	1.1	41.7	55.4	73.5	77.9	88.9	3.1	1.7	0	0	0
		20	3.1	1.9	0	17.2	0	2.5	0	0	8.9	0	58.4	76.3	83	84.6	100	0	0	0	0	0
	EF: 0.5	4	5.4	8.4	14.6	20	22.9	2.4	7.3	10.5	11.6	15.8	58.8	75.9	84	90.5	90.9	1.2	1.6	5.1	9.4	11.4
		8	6	22.8	48	56	79.6	5.3	14.6	23.2	30.1	52.8	54.2	77.7	87.7	89.2	97.2	2.2	7.2	12.1	22.3	32.5
		10	28	42.5	50.7	80.5	86.8	21.1	30.1	31.9	56.6	69.4	70.7	85.4	89	96	94.3	1.4	14.2	16.5	34.7	57.9
		14	30.2	54.1	83.4	98.6	100	22.8	34.1	68.4	86.2	96.3	80.1	89.6	98.5	100	100	7.7	19.1	42.2	56.8	83.5
		20	42.5	100	100	100	100	33.7	86.2	100	100	100	91.7	96.7	98.8	99	100	6.2	37.8	80.1	95.9	100
	EF: 0.8	4	14.4	32.9	42.7	65.7	81.3	9.7	25.4	36.3	48.5	56.8	78.1	94.6	98.7	100	100	3.5	7.1	18.3	31.1	38.7
		8	36.5	69.8	95.7	98.4	100	22.4	59.2	78.5	93.5	100	79.4	93.8	94.7	100	100	5.8	25.7	53.4	75.1	92.4
		10	55.1	82.6	100	100	100	49.4	71	93.7	100	100	86.6	98.2	99.9	100	100	24.2	46.5	75.6	96	94.2
		14	66.3	98.6	100	100	100	54.7	88	99.3	100	100	99.9	99.9	99.9	100	100	24.2	62.7	95.6	100	100
		20	97.4	100	100	100	100	94.8	100	100	100	100	99.1	99.9	100	100	100	44.3	100	100	100	100
	EF: 1.2	4	32	72.6	95.8	99	97.6	26.2	64.4	90	94.7	95.4	94	99.9	99.9	100	100	5	26.4	61.6	80.8	88
		8	74.6	100	100	100	100	55.6	96.6	100	100	100	95.1	99.3	99.9	100	100	17.8	76.6	96.8	100	100
		10	78.9	100	100	100	100	75	100	100	100	100	99.9	99.9	99.9	100	100	45.6	84.9	100	100	100
		14	100	100	100	100	100	100	100	100	100	100	99.9	99.9	99.9	100	100	67.5	100	100	100	100
		20	100	100	100	100	100	100	100	100	100	100	99.9	99.9	100	100	100	98.1	100	100	100	100