Procedimientos de medición del tamaño funcional: un mapeo sistemático de literatura

Christian Quesada-López, Marcelo Jenkins

CITIC, Universidad de Costa Rica, San Pedro, Costa Rica
{cristian.quesadalopez, marcelo.jenkins}@ucr.ac.cr

Resumen. La medición del tamaño funcional (FSM) del *software* ha sido ampliamente reconocida por su utilidad en distintas fases de desarrollo y por su independencia de la tecnología. Una de las principales limitantes para la adopción de los métodos FSM es el tiempo y el costo del proceso de conteo y la confiabilidad de los resultados. La formalización de los procedimientos de FSM incrementa las posibilidades de automatización porque reducen la variabilidad que genera la interpretación de las reglas generales de los métodos FSM. En este estudio se identifican y caracterizan procedimientos de medición de tamaño funcional mediante un mapeo sistemático de literatura, los estudios se mapean de acuerdo con el método de medición, los artefactos de entrada, el método de investigación, el contexto, su nivel de automatización y las amenazas a la validez reportadas. En total 150 artículos son analizados. Existe una necesidad de evidencia empírica sobre la automatización de procedimientos FSM. Esto incluye el uso estandarizado de protocolos de evaluación y un mayor nivel de detalle en los reportes para facilitar las comparaciones y replicaciones de estudios.

Palabras clave. Tamaño funcional, procedimientos de medición, automatización, estudio de mapeo, revisión sistemática.

1 Introducción

La medición del tamaño del *software* es fundamental en la ingeniería del *software* [1]. La medición del tamaño funcional (FSM) ha demostrado su utilidad en distintas fases de desarrollo, por su independencia de la tecnología [2]. Esta es utilizada para estimar el esfuerzo y el costo de los proyectos, evaluaciones de calidad, monitoreo y control, negociación de cambios, medir la productividad y la densidad de defectos [3]. En la actualidad existen los métodos estandarizados COSMIC FFP (ISO/IEC 19761), IFPUG FPA (ISO/IEC 20926), MkII (ISO/IEC 20968), NESMA (ISO/IEC 24570) y FiSMA (ISO/IEC 29881).

Una de las principales limitantes para la adopción de métodos FSM es el tiempo y el costo del proceso de conteo [4]. Con la formalización y automatización del proceso se alcanzan varios beneficios tales como la reducción del tiempo, esfuerzo y costo, y se mejora la confiabilidad, exactitud, y repetibilidad de los conteos [5]. La automatización del proceso de conteo proporcionando mediciones confiables es un reto en la industria [6] y aunque es percibida como importante para la toma de decisiones, aún presenta

grandes variaciones en la exactitud de los resultados, que no son aceptables para los profesionales [5]. La dificultad para evaluar las propuestas de medición en un nivel práctico se da por la falta de rigurosidad en las validaciones empíricas; además, los resultados de medición tienen que analizarse detalladamente, ya que todos los métodos FSM tienen sus limitaciones y muchos de los resultados realizados mediante procesos rápidos y automatizados, no cuentan con un respaldo científico comprobado [7]. Para que las mediciones de tamaño funcional sean consideradas una métrica, deben estar definidas con precisión, tener propiedades matemáticas claras, y ser demostrable en términos de precisión y exactitud, validez y confiabilidad [1].

Un procedimiento FSM presenta un conjunto de operaciones, descritas específicamente, para obtener una medición conforme a un método FSM específico de manera sistemática y repetible [1]. Cada procedimiento utiliza un conjunto de artefactos de *software* como entrada para llevar a cabo el proceso de medición aplicando un conjunto de reglas de mapeo para identificar los componentes funcionales [8]. La formalización de estos procedimientos incrementa las posibilidades de automatización porque resuelven las ambigüedades de los métodos existentes, sus conceptos y sus reglas de aplicación para artefactos específicos, y reducen la variabilidad que genera la interpretación de las reglas generales [9]. La evidencia empírica sobre estos procedimientos es necesaria para incrementar su aplicación en la industria [5] [10]. Para demostrar la exactitud de los procedimientos FSM, es necesario definir y utilizar protocolos de verificación de exactitud estandarizados que aseguren que el proceso de medición produzca los resultados correctos. En la actualidad existe poca investigación acerca de la verificación de los resultados producidos por los diferentes procedimientos. Son requeridos procedimientos formales, automatizados y validados que puedan ser auditados para garantizar la calidad de los resultados del proceso de medición [11].

El objetivo de este estudio es identificar y caracterizar literatura existente sobre los procedimientos de medición del tamaño funcional. El análisis incluye consideraciones para mitigar algunas de las limitaciones de la investigación en el área y la síntesis de los principales resultados obtenidos a partir de la evidencia existente. Las preguntas de investigación son: (**RQ1**) ¿Cuáles son las características de las propuestas de procedimientos de medición del tamaño funcional que han sido publicadas en la literatura? (**RQ2**) ¿Cómo han sido evaluados los procedimientos de medición del tamaño funcional? (**RQ3**) ¿Qué amenazas a la validez se han reportado en el diseño y evaluación empíricos de los procedimientos de medición del tamaño funcional? Para responderlas se realiza un mapeo sistemático de literatura que permite clasificar la evidencia existente en cuanto a los métodos de medición utilizados, los artefactos de entrada y las reglas de mapeo, el método de investigación usado para determinar su eficacia, el contexto de aplicación, el nivel de automatización alcanzado y las amenazas a la validez reportadas durante el diseño y la ejecución.

2 Estudios secundarios sobre la medición del tamaño funcional

Múltiples estudios secundarios han analizado distintos aspectos sobre la medición del tamaño funcional (FSM). Abran, Meli y Symons [12] presentan el estado del arte y

perspectivas futuras del método COSMIC. Las principales preocupaciones en la industria son el desempeño de la medición y las evaluaciones comparativas. Se requiere mejorar el rendimiento de los procesos de conteo e incrementar la adopción de programas de medición en la industria que incluyan entandares de medición. Stambollian y Abran [4] presentan un marco de trabajo de referencia para profesionales con el conjunto esencial de funciones para la implementación de estándares FSM. Incluyen una recopilación de las herramientas de medición disponibles para COSMIC y determinan que ninguna de ellas automatiza totalmente el proceso de medición. Es necesaria la evaluación de las herramientas para establecer su efectividad y la automatización debe considerar diferentes artefactos de entrada y reducir la intervención de un experto humano.

Marín, Giachetti y Pastor [8] realizan un análisis de los métodos FSM que utilizan modelos conceptuales como artefactos de entrada para la medición COSMIC. Los autores realizan recomendaciones para el desarrollo de los procedimientos FSM y determinan la necesidad de establecer claramente las fases de la estrategia de medición, las reglas de mapeo entre los conceptos del método y los artefactos de entrada, y las reglas de identificación de los elementos funcionales del método FSM. Determinan que la automatización es necesaria para reducir el costo del conteo e incrementar la eficiencia del proceso de medición.

Gencel y Demirors [13] identifican y evalúan los métodos de medición del tamaño funcional y sus oportunidades de mejora. Plantean que la investigación en el área se debe direccionar hacia la evaluación de la significancia de las reglas del proceso de medición, el análisis del nivel de abstracción de las entidades de medición y la determinación de los efectos de los niveles de granularidad de los atributos medidos. Finalmente, identifican la necesidad de desarrollar herramientas para la automatización parcial o total del proceso de conteo de tamaño funcional.

Ozkan y Demirors [9] analizan propuestas de formalización FSM. Estas buscan resolver las ambigüedades de los métodos de medición existentes, sus conceptos y sus reglas de aplicación, e intentan reducir la variabilidad que genera la interpretación de las reglas generales de estos métodos de medición. Dada la formalización de los métodos FSM a partir de la aplicación de reglas específicas a especificaciones formales del *software*, es posible explorar las posibilidades de automatización. La automatización de los procesos de medición puede ser parcial o total, y se debe realizar a partir de la interpretación consistente de las reglas de medición, para garantizar la consistencia y la repetibilidad de los resultados obtenidos por estos procedimientos. Finalmente, indican que las definiciones formales permiten comparar transparentemente las mediciones entre los diferentes modelos.

Barkallah, Gherbi, y Abran [14] proponen un marco de trabajo para la automatización de los procedimientos FSM COSMIC basados en modelos UML como artefactos de entrada. Determinan que existe poca evidencia sobre la validación de las herramientas existentes que proponen la automatización de la medición del tamaño funcional. Bajwa, Gencel y Abrahamsson [10] realizan un mapeo sistemático para identificar los métodos de medición del tamaño del *software* incluyendo los de tamaño funcional. Identifican las entidades medidas y el tipo de validación empírica. Discuten el estado del arte del área de la medición del tamaño del *software* y determinan que la mayoría

de los estudios proponen métodos FSM. La mayoría de estudios analizan métodos existentes de tamaño funcional, lo cual corrobora la aceptación de estas métricas en la comunidad académica. Una constante a través de los distintos estudios es la necesidad de evaluaciones empíricas sobre la efectividad de los métodos FSM. El presente mapeo sistemático confirma algunos de los hallazgos anteriores, agrega nueva evidencia y responde nuevas preguntas de investigación en el área de estudio.

3 Metodología del mapeo sistemático de literatura

A continuación se presentan los pasos del proceso del mapeo de acuerdo con los lineamientos propuestos en [15] y las recomendaciones planteadas en [16] [17]. El detalle de los instrumentos utilizados en el protocolo de mapeo se encuentra disponibles en https://goo.gl/lgfYh8.

Estrategia de búsqueda y proceso de selección. Se realizó una búsqueda exploratoria para identificar estudios relevantes que son usados como artículos de control. Esta búsqueda se basó en las preguntas de investigación, en términos utilizados en los estudios secundarios relacionados en el área y en experiencias y lecciones aprendidas durante el estudio realizado en el 2014 [18]. A partir de este proceso se selecciona 11 artículos de control y se determinan los términos de búsqueda para el estudio. La cadena de búsqueda se construye a partir de la extracción de términos clave del título y del resumen del conjunto de artículos de control. Se desarrolla el modelo PICO (Población, Intervención, Comparación, Salidas) de acuerdo con [16]. La cadena de búsqueda utilizada es: **Población** (aplicaciones de *software* y proyectos de *software*. Ejemplo: "sistema", "*software*", "aplicación"), **Intervención** (medición de tamaño funcional. Ejemplo: "puntos de función", "tamaño funcional", "medición"), **Comparación** (no fue considerada) y **Salidas** (procedimientos, reglas de mapeo, formalización o automatización de la medición del tamaño funcional. Ejemplo: "automatización", "reglas mapeo", "sistemático", "herramienta", "representación formal").

La cadena de búsqueda fue refinada mediante un conjunto de pruebas piloto para reducir el ruido. Con la cadena aprobada se realizan las búsquedas en las bases de datos digitales tal como se muestra en el Cuadro 1. Las bases de datos fueron seleccionadas de acuerdo con las recomendaciones de estudios relacionados en el área de interés e incluyen *SCOPUS, IEEE Xplore, ISI Web of Science e Engineering Village.*

El protocolo bases del mapeo fue desarrollado durante el 2014 [18] y actualizado y corrido durante el primer semestre del 2016. La búsqueda automatizada se realiza en mayo del 2016 y los estudios se analizan durante el segundo semestre del 2016. La búsqueda considera estudios primarios recuperados hasta el año 2016.

Cuadro 1 Búsqueda en bases de datos

TITLE-ABS-KEY(("software" OR "web*" OR "system*" OR "application*") AND ("function* point*" OR "functional size*" OR "function* measurement") AND ("automat*" OR "systematic" OR "procedur*" OR "tool*" OR "mapping*" OR "rule*" OR "formal * representation"))*

El Cuadro 2 muestra el número de resultados por base de datos (n), los estudios incluidos después de eliminar duplicados, los resultados del proceso de inclusión y exclusión (I/E) y finalmente, los artículos de control identificados en cada base de datos. Se utiliza Mendeley como herramienta de administración de referencias y MS Excel para los procesos de selección, evaluación y extracción de datos.

Cuadro 2 Número de estudios por bases de datos

Base de datos	n	I/E	Control	Base de datos	n	I/E	Control
SCOPUS	722	97	14	*ISI Web of Science*	165	21	5
IEEE Xplore	81	29	3	*Engineering Village*	128	53	10

El proceso de I/E se realiza basado en los títulos, resúmenes de los artículos y ante el caso de duda, en la lectura del texto completo. Adicionalmente, se identificaron estudios a través del proceso de "bola de nieve hacia atrás" y "hacia adelante" basado en los artículos relevantes de la búsqueda automática y siguiendo las recomendaciones descritas en [19]. El proceso se realiza de la siguiente manera: (1) un investigador realiza la lectura e incluye o excluye los artículos basado en el título, resumen o lectura completa, (2) un segundo investigador, que actúa como consultor, evalúa una muestra aleatoria de los artículos y valida los resultados. Para este proceso se definen criterios específicos de inclusión y exclusión.

En nuestro estudio excluimos publicaciones que no cumplen con la fórmula (E1 OR E2 OR E3 OR E4) donde: (**E1**) Estudios secundarios o terciarios. (**E2**) Estudios en idiomas diferentes al inglés. (**E3**) Estudios relacionados con la estimación de tamaño funcional a partir de métodos de conversión y/o mediante técnicas de minería o inteligencia artificial sin considerar el proceso de medición. (**E4**) Estudios que traten únicamente el tema de modelos de estimación basados en el tamaño funcional del *software* sin considerar el proceso de medición.

Los estudios que no fueron excluidos y que cumplen con la fórmula (I1 AND I2 AND (I3 OR I4)) son incluidos de acuerdo con: (**I1**) Estudios relacionados con el área de medición del tamaño funcional del *software* en el campo de la ingeniería del *software*. (**I2**) Estudios primarios relacionados con un procedimiento de medición de tamaño funcional basado en los métodos de tamaño funcional estandarizados (IFPUG FPA, COSMIC FFP, MARK II, FiSMA y NESMA). (**I3**) Estudios que proveen alguna descripción del procedimiento de medición del tamaño funcional. (**I4**) Estudios primarios que proveen algún detalle de la evaluación de los procedimientos de medición del tamaño funcional. En este proceso se aplica una validación *test–retest* [20].El criterio inclusivo permite seleccionar procedimientos de medición que realizan modificaciones a los métodos estandarizados y procedimientos de estimación de tamaño funcional.

En total, 150 artículos son identificados y analizados. En la búsqueda en bases de datos digitales se identifican 102 artículos que incluyen los 11 artículos de control y como resultado del proceso de bola de nieve se identifican 48 artículos relevantes: 28 hacia atrás y 20 hacia adelante, tal como se muestra en la Figura 1.

Evaluación de calidad. La evaluación de la calidad refleja el nivel de detalle ofrecido para los aspectos más relevantes del análisis. Los artículos seleccionados son evaluados

de acuerdo al siguiente conjunto de criterios de calidad: (**Q1**) ¿El artículo menciona claramente el método de medición utilizado? (**Q2**) ¿El artículo describe claramente o referencia los artefactos de entrada? (**Q3**) ¿El artículo explica o referencia las reglas de mapeo del procedimiento? (**Q4**) ¿El artículo menciona el desarrollo de una herramienta de automatización? (**Q5**) ¿El artículo describe el proceso experimental de validación? (**Q6**) ¿El artículo describe las amenazas a la validez y sus respectivas acciones? (**Q7**) ¿El artículo presenta los resultados de la validación empírica? Los valores de la evaluación varían entre 0 y 17. Un puntaje de calidad mayor indica que el reporte provee un mayor nivel de detalle y no debe asociarse con la calidad del estudio, ni la calidad del foro de publicación. En general, los estudios requieren mayor nivel de detalles relacionados con la experimentación (54%) y las amenazas a la validez (83%). La mediana de las evaluaciones es 11, el Q25 es 9 y el Q75 es 13.

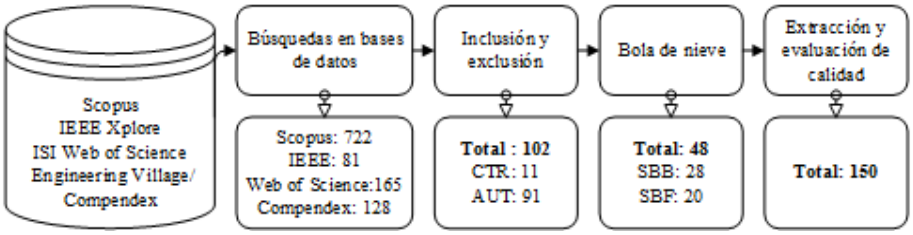

Figura 1 Artículos incluidos durante el proceso de selección

Extracción de datos. Para los artículos identificados se selecciona la información relevante para el análisis. Los componentes del formulario de extracción asociados a cada pregunta de investigación incluyen: (RQ1) Información general del mapeo con el título, año, autores, foro de publicación, enfoque, método, artefacto de entrada, reglas de mapeo, dominio funcional, tipo de artículo y nivel de automatización. (RQ2) método de investigación, tipo de evaluación, participantes y métricas. (RQ3) Amenazas a la validez. La extracción fue realizada por un investigador y validada por un segundo investigador, que actúa como consultor, a partir de una muestra aleatoria de los artículos extraídos.

Análisis y clasificación. Para el análisis y síntesis de la información se aplican las siguientes estrategias. Para el análisis cuantitativo de las preguntas de investigación RQ1 y RQ2, se construye el esquema de clasificación de acuerdo con el procedimiento en [21], análisis temático de acuerdo con [22] y una validación de las categorías de acuerdo con [8]. Para el análisis cualitativo de la pregunta RQ3, se utiliza el análisis narrativo que resume y describe los hallazgos y la evidencia extraída [15] [22]. Los artículos relacionados a una sola propuesta son contabilizados por separado.

Amenazas a la validez. Identificación de estudios primarios. La cadena de búsqueda fue definida a partir de un conjunto de artículos de control, y piloteada en varias pruebas, para reducir el ruido. Las bases de datos seleccionadas son reconocidas por tener gran cobertura en el área de la ingeniería de *software* (SE). La búsqueda automática se complementa con el proceso de bola de nieve. Ante dudas sobre la inclusión de un

artículo se realiza la lectura completa. Se excluye literatura gris y artículos que no están en inglés. Los estudios secundarios identificados durante el proceso validan la selección de artículos primarios. **Extracción y clasificación.** La extracción se basa en un esquema de clasificación específico. El proceso de clasificación y extracción es realizado por un solo investigador y validado por un segundo investigador, se ejecutan revisiones ciegas aleatorias en un subconjunto de artículos. Las ambigüedades se resuelven durante el proceso. La interpretación de los resultados por parte de los investigadores es una amenaza a la validez. Los artículos son clasificados de acuerdo con lo reportado por los autores originales y en caso de no ser reportados explícitamente los investigadores de este estudio, si es posible, realizan la clasificación. Se realiza un proceso sistemático para la clasificación de los artículos. Se diseña un formulario de extracción para la recolección de datos que guía el proceso y que puede ser revisado. La aplicación de los criterios de calidad es realizada solo por un investigador, lo que representa una amenaza a la validez. **Generalización de los resultados.** La generalización de resultados se limita a los estudios incluidos en el mapeo. Los investigadores tienen experiencia en el área de medición, lo que es una ventaja, pero a su vez, cabe la posibilidad de que sus intereses y sus opiniones afecten las recomendaciones que se realizan. Durante todo el proceso, se aplican lineamientos en el área de la SE experimental que detallan los protocolos y las mejores prácticas para ejecutar estudios secundarios. Se reporta detalladamente el proceso para facilitar el análisis y replicación.

4 Resultados

El listado completo de publicaciones resultado del mapeo se encuentra disponibles en https://goo.gl/lgfYh8, ordenado por orden cronológico, empezando por el más reciente. Los artículos son identificados con el prefijo [S1-S150]. La lista detalla la caracterización del procedimiento FSM, la validación empírica realizada, las limitaciones presentadas y la evaluación de calidad. La lista de artículos correspondientes para cada una de las clasificaciones descritas en el mapeo es detallada. A continuación se presentan los resultados del mapeo sistemático de literatura.

RQ1. Frecuencias y características de las publicaciones. ¿Cuáles son las características de las propuestas de procedimientos de medición del tamaño funcional que han sido publicadas en la literatura?

Frecuencia de publicaciones. Las Figura 2 muestra la tendencia total de publicaciones identificadas entre los años 1991-2016 donde un total de 150 artículos son identificados. Un total de 73 publicaciones se basan en (C) COSMIC (48.6%), 72 en (I) IFPUG (48.0%) y 2 en (M) Mark II (1.3%). Dos artículos presentan trabajos que integran COSMIC, IFPUG y Mark II (1.3%), y uno integra COSMIC e IFPUG (0.6%). El método COSMIC presenta una tendencia creciente constante, el método IFPUG se ha mantenido estable a través del tiempo y el método Mark II tiene pocas publicaciones. No se identificaron artículos para los demás métodos FSM. A partir del 2010 la tendencia de publicaciones para COSMIC supera las del método IFPUG. El año de mayor cantidad

de publicaciones para COSMIC es el 2010 (10), seguido del 2011 (9), 2013 (8) y 2015 (8); para IFPUG es el 2005 (6), 2006 (6) y 2008 (6).

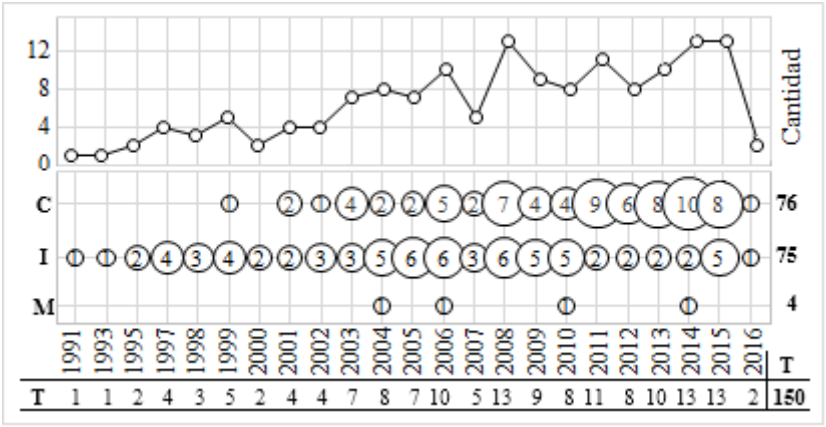

Figura 2 Publicaciones por año y por método FSM

Foros de publicación. En total, 112 artículos fueron publicados en conferencias (75%), 33 en revistas (22%), 2 artículos son capítulos de libros (1%) y 3 reportes técnicos (2%). Los capítulos y reportes técnicos son considerados por pertenecer a autores reconocidos en el área. El Cuadro 3 detalla el *top* 4 de las (C) conferencias y las (J) revistas donde los artículos del mapeo fueron publicados. La principal conferencia fue *Software Measurement and Software Process and Product Measurement* (IWSM-MENSURA) con 19 artículos basados en (C) COSMIC y 6 artículos basados en (I) IFPUG, seguida de *Empirical Soft. Engineering and Measurement / Software Metrics* con (METRICS & ESEM) 1 artículo basados en COSMIC y 8 artículos basados en IFPUG. La principal revista fue *Information and Software Technology* (IST) con 4 artículos basados en COSMIC y 2 artículos basados en IFPUG.

Artefactos de entrada. Los procedimientos FSM aplican un conjunto de reglas de mapeo entre componentes de los métodos FSM y los conceptos de los artefactos de *software* utilizados para la medición. La Figura 3 detalla la tendencia de los artefactos usados en las publicaciones de los procedimientos FSM. En total, 136 artículos utilizan artefactos desarrolladas en las (E) etapas tempranas del desarrollo del *software* (88%), 19 artículos del (D) código fuente (12%).

Cuadro 3 Foros de publicación

T	R	Foro	n	%	C	I	M	T	R	Foro	n	%	C	I	M
C	1	IWSM-MENSURA	25	23	19	6	0	C	4	SMEF	4	4	1	4	0
C	2	METRICS & ESEM	9	8	1	8	0	*Artículos en revistas*							
C	3	QSIC	5	5	4	1	0	J	1	IST	6	18	4	2	0
C	3	SEAA	5	5	4	1	0	J	2	JSS	3	9	1	2	1
C	4	CibSE	4	4	1	2	1	J	3	ESE	2	6		2	0
C	4	PROFES	4	4	3	1	0	J	4	TOSEM	2	6	1	2	0

En etapas tempranas, un total de 70 publicaciones se basan en (C) COSMIC (51%), 62 en (I) IFPUG (46%) y 4 en (M) Mark II (3%). El método COSMIC presenta la mayor cantidad de publicaciones en el 2011 (9) y 2014 (9), el método IFPUG se mantuvo estable entre el 2004-2010 con 5 publicaciones en promedio. De los artículos que utilizan el código fuente, 13 publicaciones se basan en IFPUG (68%) y 6 en COSMIC (32%). El método COSMIC presenta la mayor cantidad de publicaciones en el 2011 (9) y 2014 (9), el método IFPUG se mantuvo estable entre el 2004-2010 con 5 publicaciones en promedio. El año de mayor cantidad de publicaciones para COSMIC es el 2015 (2) y relacionadas con IFPUG es el 2015 (4), principalmente relacionados con el análisis de OMG AFP [28].

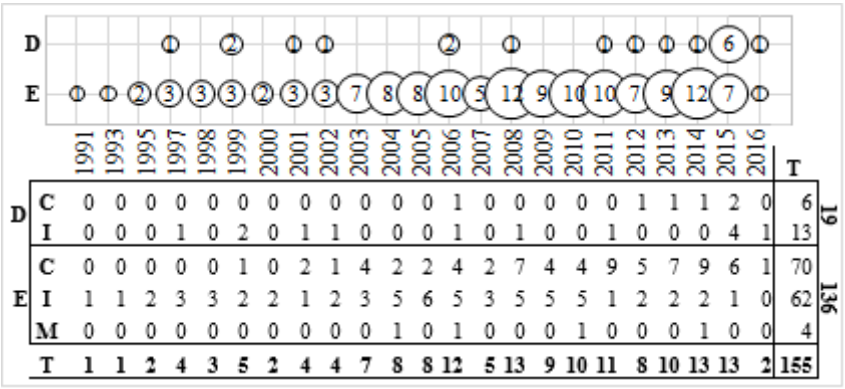

Figura 3 Artefactos por año y por método FSM

El 41% de las publicaciones utilizan artefactos basados en modelos UML (63) tales como: diagramas de clase, objetos, secuencia, actividad, casos de uso y componentes; 34% modelos conceptuales (52) que utilizan modelos de clases, dinámicos, flujo de datos, funcionales y de presentación; 5% especificaciones de requerimientos (8) como requerimientos funcionales, historias de usuarios y ontologías de requerimientos, 4% lenguajes formales de especificación (6) como metas y escenarios, notaciones matemáticas y especificaciones VDM-SL, 3% modelos de negocio (4) como BPMN y 2% modelos de diseño particulares (3) tales como modelos propietarios, de datos y de interfaz. Las líneas de investigación que se han mantenido activas a través de los años son las que utilizan modelos UML (33 IFPUG, 28 COSMIC, 2 Mark II) y modelos conceptuales (21 IFPUG, 30 COSMIC, 1 Mark II). Las propuestas para calcular el tamaño funcional a partir de modelos de negocio (BPM) solo se han realizado para COSMIC, entre el 2011-2014.

Caracterización. Las publicaciones desarrollan propuestas, proponen herramientas para automatizar el proceso de medición y/o validan empíricamente los resultados. Un total de 49 artículos realizan una propuesta, una herramienta y la validación (33%), 45 una propuesta y herramienta (30%), 25 una propuesta y validación (17%), 24 solo una

propuesta (16%), que incluye un análisis crítico del AFP y 10 realizan solo una validación empírica (7%) basado en un trabajo previo y de los cuales 6 utilizan una herramienta. El 60% de las publicaciones dicen ser de medición de tamaño funcional (90) y el 40% de estimación de tamaño funcional (60). Todos los artículos miden la entidad producto. Un total de 90 (60%) publicaciones son para ambientes y metodologías orientadas a objetos (OO), 34 (23%) para desarrollo dirigido por modelos (MDA), 17 para aplicaciones web (11%), 5 publicaciones para aplicaciones móviles (3%) y solo se identifica una publicación para ambientes ágiles (1%). Un total de 22 (15%) publicaciones son para el dominio funcional de sistemas de tiempo real (RTS) y 127 (85%) para el dominio de sistemas de información (MIS). IFPUG es utilizado en MIS (47%) y en RTS (1%), COSMIC en MIS (36%) y en RTS (13%). Las publicaciones desarrollan herramientas para apoyar la automatización del proceso de conteo en diferentes niveles. El 83% (124) de las publicaciones mencionan la posibilidad de automatización a partir del procedimiento FSM y el 65% (98) utilizan una herramienta. La mayoría de las herramientas realizan análisis estático sobre los afectos de *software*. El análisis dinámico se realiza sobre artefactos de *software* como el código fuente y las aplicaciones corriendo en ambientes controlados. El 35% utilizan herramientas para IFPUG y el 31% para COSMIC.

RQ2. Evaluación empírica de los procedimientos. ¿Cómo han sido evaluados los procedimientos de medición del tamaño funcional?

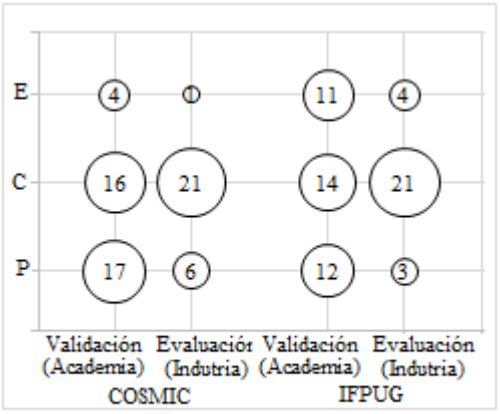

Figura 4 Evaluación empírica

Evaluación. El 87% de las publicaciones (130) realizan algún tipo de validación para los procedimientos, tal como se muestra en la Figura 4. En total, 74 publicaciones realizan validaciones en ambientes académicos (55%) y 56 evaluaciones en la industria (44%). Las validaciones en la industria varían entre la industria automotriz, financiera, gobierno, telecomunicaciones, juegos, entre otras. Las evaluaciones realizadas en la industria para el método COSMIC representan el 22% y las validaciones en la academia el 26%. El mismo comportamiento se presenta para IFPUG. El 15% realizan procesos (E) experimentales (20), el 55% (C) casos de estudio (72) y el 29% (P) pruebas de concepto o estudios piloto (38). El método COSMIC ha sido evaluado principalmente mediante casos de estudio (28%) y pruebas de concepto (18%), IFPUG mediante casos de estudio (26%) y pruebas de concepto (12%) del total de publicaciones. En las evaluaciones en la industria, el método principal es el de caso de estudio, con un 75% de las veces. En las validaciones en la academia, el 40% son pruebas de concepto, 39% caso de estudio y el 21% procesos experimentales. Los estudios experimentales regularmente realizan

procesos con estudiantes para las validaciones (80%). Solo cinco publicaciones se reportan como replicaciones de estudios anteriores.

Criterios de evaluación. Los estudios validan las mediciones comparando la exactitud entre el valor obtenido por el procedimiento y un valor de referencia logrado mediante la medición de uno o varios expertos (68%), validando la repetibilidad o reproducibilidad (13%), propiedades de adopción como facilidad de uso, utilidad e intención de uso futuro (10%) y productividad (10%). En cuanto a la evaluación de exactitud, se identificaron 3 protocolos de verificación: 2 para COSMIC y 1 para IFPUG, todos basados en la propuesta descrita en [13]. Solo 6 publicaciones mencionan explícitamente el uso de un protocolo de verificación de exactitud.

RQ3. Limitaciones reportadas en las publicaciones. ¿Qué amenazas a la validez se han reportado en el diseño y evaluación empíricos de los procedimientos de medición del tamaño funcional?

El 25% (38) de las publicaciones reportan amenazas a la validez y solo el 12% (19) las clasifican en un tipo específico. Las principales limitaciones se relacionan con la dependencia de los contextos de desarrollo. Para aplicar los procedimientos son necesarios artefactos específicos, y en muchos casos, se evalúan con aplicaciones pequeñas y en ambientes académicos. Algunos artículos solo realizan pruebas de concepto para la validación. En la mayoría de los casos, el procedimiento se limita para un solo método, lo cual no permite la comparación de resultados entre métodos; finalmente, la mayoría de los procedimientos no utilizan un protocolo sistematizado para evaluar los resultados. A continuación se reportan las principales amenazas reportadas por los autores.

Validez interna. La utilización de personas (estudiantes, profesionales y expertos) presenta una variación natural en la aplicación de los procesos de medición. Los experimentos en la academia con estudiantes seleccionados a conveniencia se reporta como una amenaza, pero se reporta que el nivel de conocimiento y experiencia similar permite comparar los rendimientos obtenidos. Se recomienda validar la confiablidad en la recolección de los datos y mantener la motivación de los participantes; además, seleccionar aleatoriamente a los participantes. El uso de un solo experto para determinar las medidas de referencia es una amenaza a la validez; algunos estudios indican la participación de varios expertos para determinar ese valor de referencia y, en algunos casos, los valores de referencia no son calculados por contadores certificados. El tamaño y la cantidad de las aplicaciones y artefactos utilizados en los estudios representan una amenaza, el contexto y nivel de detalle de los artefactos de *software* para la aplicación de los procedimientos son un factor de influencia en los resultados. Los estudios que comparan procedimientos fijan este factor y utilizan artefactos del mismo dominio funcional y con características funcionales similares. Se recomienda el uso de artefactos estandarizados de acuerdo a especificaciones reconocidas y la utilización de material revisado y validado en estudios anteriores. **Validez del constructo.** Los estudios reportan métricas de evaluación ampliamente aceptadas en el campo de medición del tamaño funcional. En el caso de propiedades de efectividad se utilizan métricas reconocidas, como

exactitud, repetibilidad, reproducibilidad y productividad. Respecto a las propiedades de adopción, se utilizan modelos de adopción de tecnologías basados en [30] y adaptados para métodos de medición. Algunos investigadores advierten que las herramientas automatizadas pueden no seguir completamente las reglas establecidas por los estándares de medición. **Validez de las conclusiones.** Los autores reportan la heterogeneidad de los participantes, aplicaciones y artefactos utilizados como una amenaza; además, de la cantidad y tamaño limitado de los artefactos utilizados para la experimentación. **Validez externa.** El uso de aplicaciones pequeñas en contextos académicos es una limitación. Se deben utilizar aplicaciones similares a las que se encuentran en la industria o aplicaciones reales de la industria. La generalización de los resultados se ve limitada porque los procesos de conteo se aplican en contextos específicos de desarrollo, bajo un dominio funcional particular y con artefactos de *software* específicos. Los estudios reportan que la participación de sujetos que no son regularmente usuarios de los procedimientos de medición, representan una amenaza para generalizar los resultados sobre el uso y la adopción.

5 Discusión

La formalización de los procedimientos FSM puede potenciar la automatización, mejorar la productividad, reducir los costos y mejorar la consistencia de los resultados. La evidencia empírica sobre cada una de las propuestas es esencial para entender su nivel de madurez. Este estudio identifica grupos de artículos relacionados entre sí, dada las características de la propuesta que desarrollan o evalúan. Por ejemplo, el Grupo 7 (**G7**) presenta 10 artículos que miden aplicaciones desarrolladas bajo el método OO-Method basado en IFPUG, utilizando como entrada modelos conceptuales. Proponen un procedimiento FSM y las reglas de mapeo. Evalúan los resultados de acuerdo a la exactitud, reproducibilidad, productividad y propiedades de adopción. Finalmente, aplican los resultados en modelos de estimación de esfuerzo. El **G6** presenta 8 artículos que detallan el procedimiento *OOmFP* para medir automáticamente aplicaciones desarrolladas con *OO-Method*, basado en COSMIC; propone el procedimiento FSM y sus reglas de mapeo; evalúan los resultados de acuerdo a la precisión y productividad; finalmente, aplican los resultados para la detección de defectos en modelos conceptuales. El **G9** presenta 7 artículos que detallan un procedimiento de medición para sistemas orientados a objetos y utiliza el método COSMIC sobre las especificaciones del método *OO-Method*. El **G26** presenta 5 artículos que automatizan la medición COSMIC a partir de las especificaciones de la herramienta Simulink y validan el trabajo realizado en distintos ambientes de desarrollo en la industria. Además proponen un procedimiento de validación de exactitud sistemático.

Finalmente, otros grupos presentan procedimientos para la medición de aplicaciones móviles con COSMIC (G0), analizan propuestas de automatización a partir del código fuente (G1, G5, G24), a partir de modelos UML (G2, G4, G8, G11, G13, G15, G16, G17, G18, G20, G21, G22, G23, G25) y modelos conceptuales (G10, G12, G14, G26, G27. G29), especificaciones con metas y escenarios (G3), ontologías de requerimientos (G28) y modelos de negocio (G19).

La medición del tamaño funcional sobre modelos conceptuales y UML son las más estudiadas, existe evidencia empírica sobre su aplicabilidad y se han propuesto herramientas para su automatización. Sin embargo, el nivel de granularidad con que se detallan los artefactos de entrada impacta la exactitud de las mediciones.

Las herramientas de automatización de medición deben ser validadas mediante protocolos sistemáticos para garantizar la confiabilidad de los resultados. No solo es necesario evaluar los resultados totales de medición, sino también los resultados obtenidos para cada uno de los componentes funcionales. Se deben formalizar marcos de trabajo para la automatización del conteo de puntos de función y validarlos mediante el uso de protocolos de verificación, como los propuestos en S1, S2, S13, S27 y S84.

En general, existen pocas replicaciones para las diferentes propuestas de los procedimientos que permitan validar y generalizar los resultados. En particular, no se encontraron herramientas académicas disponibles en línea para realizar experimentación sobre procedimientos de medición específicos, lo que limita la posibilidad de realizar replicaciones. Algunos estudios han generado instrumentos y artefactos de referencia para utilizar en las validaciones de las nuevas propuestas de procedimientos FSM los cuales los investigadores deben aprovechar.

6 Conclusiones y trabajo futuro

Existe una necesidad de evidencia empírica sobre la automatización de procedimientos FSM. Esto incluye el uso estandarizado de protocolos de evaluación y un mayor nivel de detalle en los reportes para facilitar las comparaciones y replicaciones de estudios. La formalización de los artefactos de entrada, así como reglas de mapeo permite un mayor nivel de automatización; sin embargo, las diferencias de los objetos utilizados en las evaluaciones empíricas limitan la comparación de los resultados. Aunque la información sobre la eficacia y eficiencia de las propuestas se encuentran en la mayoría de los estudios, es difícil determinar todos los factores de influencia para realizar una comparación entre ellos. Existe una necesidad de procedimientos automatizados para diferentes dominios funcionales y diferentes modelos de desarrollo de *software*. La formalización en la definición de los artefactos de entrada y la rigurosa definición de las reglas de mapeo pueden mejorar el nivel de automatización. La automatización puede incrementar la productividad del proceso de medición y reducir los costos, y mejorar la consistencia.

Como trabajo futuro se pretende analizar la relación entre los artefactos de *software* reportados y la confiabilidad y exactitud de las mediciones. Interesa estudiar, cuáles son los factores que influencian los resultados de medición, la efectividad de las herramientas automatizadas, el nivel de automatización sin la intervención de un experto y el cumplimiento de cada una de las fases que los métodos estandarizados de medición establecen. Además, a partir de los resultados se debe analizar la posibilidad de establecer marcos de trabajo y de validación empírica. Finalmente, es de interés analizar los diseños experimentales y la confiabilidad de los resultados alcanzados, se debe analizar la posibilidad de sintetizar los resultados a partir de los diferentes estudios que reportan sobre una misma propuesta de medición.

Agradecimientos. Este estudio fue apoyado por el Ministerio de Ciencia, Tecnología y Telecomunicaciones (MICITT) y la Universidad de Costa Rica (No. 834-B5-A18). Agradecemos al Grupo ESE de la Universidad Federal de Río de Janeiro y al Grupo de SE de la Universidad de Costa Rica.

Referencias

1. Abran, A.: Software Metrics and Software Metrology. John Wiley & Sons, New Jersey (2010).
2. Ozkan, B., Demirors, O.: On the Seven Misconceptions about Functional Size Measurement. En: IWSM-MENSURA. pp. 45-52. IEEE (2016).
3. Garmus, D., Herron, D.: Function point analysis: measurement practices for successful software projects. Addison-Wesley Longman Publishing Co., Inc, United States (2001).
4. Stambollian, A., Abran, A.: Survey of Automation Tools Supporting COSMIC-FFP ISO 19761. COSMIC Function Points: Theory and Advanced Practices, pp: 299-316. CRC Press, Florida (2006).
5. Lavazza, L.: Automated function points: Critical evaluation and discussion. En: International Workshop on Emerging Trends in Software Metrics, WETSoM. Vol. 2015-Augus, pp. 35-43 (2015).
6. Silva, A., Pinheiro, P., Albuquerque, A.: A Brief Analysis of Reported Problems in the Use of Function Points. En: Software Engineering Perspectives Application. pp. 117-126. Springer (2016).
7. Symons, C. Lies, damned lies and software metrics. En IWSM-MENSURA, pp. 174-175 (2014).
8. Marin, B., Giachetti, G., Pastor, O.: Measurement of Functional Size in Conceptual Models: A Survey of Measurement Procedures Based on COSMIC.En: IWSM-MENSURA. pp. 170-183 (2008).
9. Ozkan, B., Demirors, O.: Formalization Studies in Functional Size Measurement. Modern Software Engineering Concepts and Practices: Advanced Approaches, 189-198 (2010).
10. Bajwa, S., Gencel, C., Abrahamsson, P.: Software Product Size Measurement Methods: A Systematic Mapping Study. En: IWSM-MENSURA. pp. 176-190 (2014).
11. Soubra, H., Abran, A., Ramdane-Cherif, A.: Verifying the accuracy of automation tools for the measurement of software with COSMIC-ISO 19761. En: IWSM-MENSURA, pp. 23-31 (2014).
12. Abran, A., Meli, R., Symons, C.: COSMIC-FFP (ISO 19761) Software size measurement: State of the art 2004. En: Software Measurement European (2004).
13. Gencel, C., Demirors, O.: Functional size measurement revisited. ACM Transactions on Software Engineering and Methodology, 17(3), 1-36 (2008).
14. Barkallah, S., Gherbi, A., Abran, A.: COSMIC functional size measurement using UML models. En: Software Engineering, Business Continuity, and Education. pp. 137-146. Springer (2011).
15. Petersen, K., Vakkalanka, S., Kuzniarz, L.: Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology, 64, 1-18 (2015).
16. Kitchenham, B., Charters, S.: Guidelines for Performing Systematic Literature Reviews in Software Engineering. Technical report, EBSE Technical Report EBSE-2007-01 (2007).
17. Biolchini, J., Mian, P., Natali, A., Travassos, G.: Systematic Review in Software Engineering. Technical Report ES 679/05, 679(May), 45 (2005).
18. Quesada-López, C., Jenkins, M., Travassos G.: Automatización de la medición del tamaño funcional del software. Reporte técnico proyecto No. 834-B5- A18, Universidad de Costa Rica (2015).
19. Wohlin, C.: Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. En: EASE pp. 1-10 (2014).
20. Kitchenham, B.: What is up with software metrics?–A preliminary mapping study. Journal of systems and software, 83(1), 37-51 (2010).
21. Petersen, K.: Measuring and predicting software productivity: A systematic map and review. Information and Software Technology, 53(4), 317-343 (2011).
22. Dixon-woods, M., Agarwal, S., Jones, D., Young, B., Sutton, A.: Synthesising qualitative and quantitative evidence: a review of possible methods. Health service & policy, 10(1), 45-53B (2005).