Universidad Nacional Autónoma de México

UNAM

Facultad de Ciencias

Integrations distinct sources databases to improve the estimation models

October 14th, 2021, Virtual IT CONFIDENCE CONFERENCE

ISBSG

IT Confidence 2021

Dr. Francisco Valdés-Souto

Associate Professor
Mathematics Department of
Science Faculty at
National Autonomous University of Mexico (UNAM)
fvaldes@Ciencias.unam.mx

COSMIC President

Mexican Software Metrics Association (AMMS), Founder

COSMIC

AMMS

Facultad de Ciencias

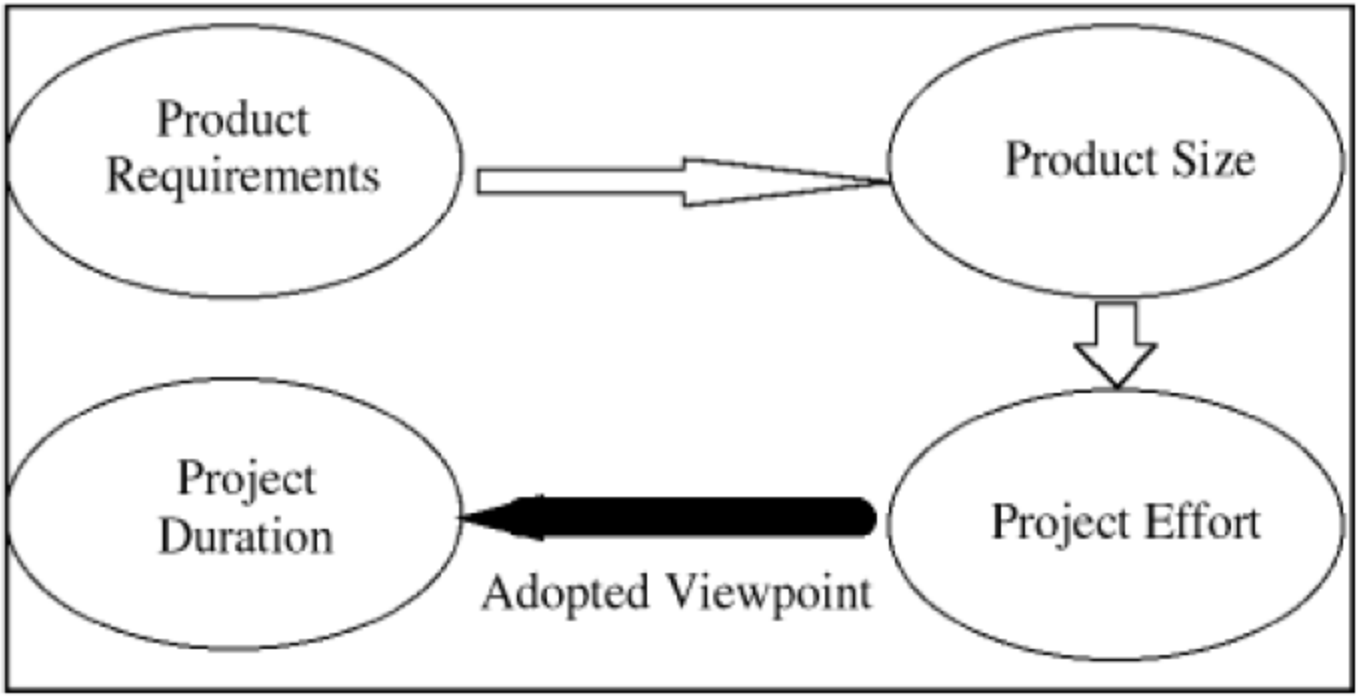
AGENDA

01 Problems with datasets in industry

02 Case study explanation

03 3 Case study results

Estimation Basics



Bourque, 2007

Several authors identify the measurement of the size of the piece of software as a relevant factor in the precision of the estimate (Linda, 2006) (Koch,2009) (De Lucia, 2005) (Hill, 2000)

Facultad de Ciencias

Problems with datasets

- Morgenshtern pointed out that "Algorithmic models **need historic data**, and many organizations do not have this information. Additionally, collecting such data may be both **expensive and time consuming.**" (Morgenshtern, 2007).

- The majority of the estimation models developed are dependent on the **representativeness of the samples** (databases) used.

- Abran mentions that "most of the so-called estimation models in the literature are productivity models." They **represent the past behavior for a specific organization** to develop software projects, representing the relationship across the two variables, usually the functional size as an independent variable and the effort or cost as dependent variables. (Abran 2015)

- In order to generate estimation models, **the researchers** have used databases documented on the **basis of the past completed projects they participated in**, usually, this **information is not available** to all the persons or is **difficult to acquire** or has elements that **do not make sense** for all the database's users.

Facultad de Ciencias

Problems with datasets

- Jørgensen et al. [1], in a systematic review of estimation studies, found "that there are good reasons to claim that the **availability of a data set is more indicative for its use than its representativeness or other properties**".

- Braga et al. [3] mention they do not found "**any reliable information about the way in which the projects included in a dataset were obtained**,"

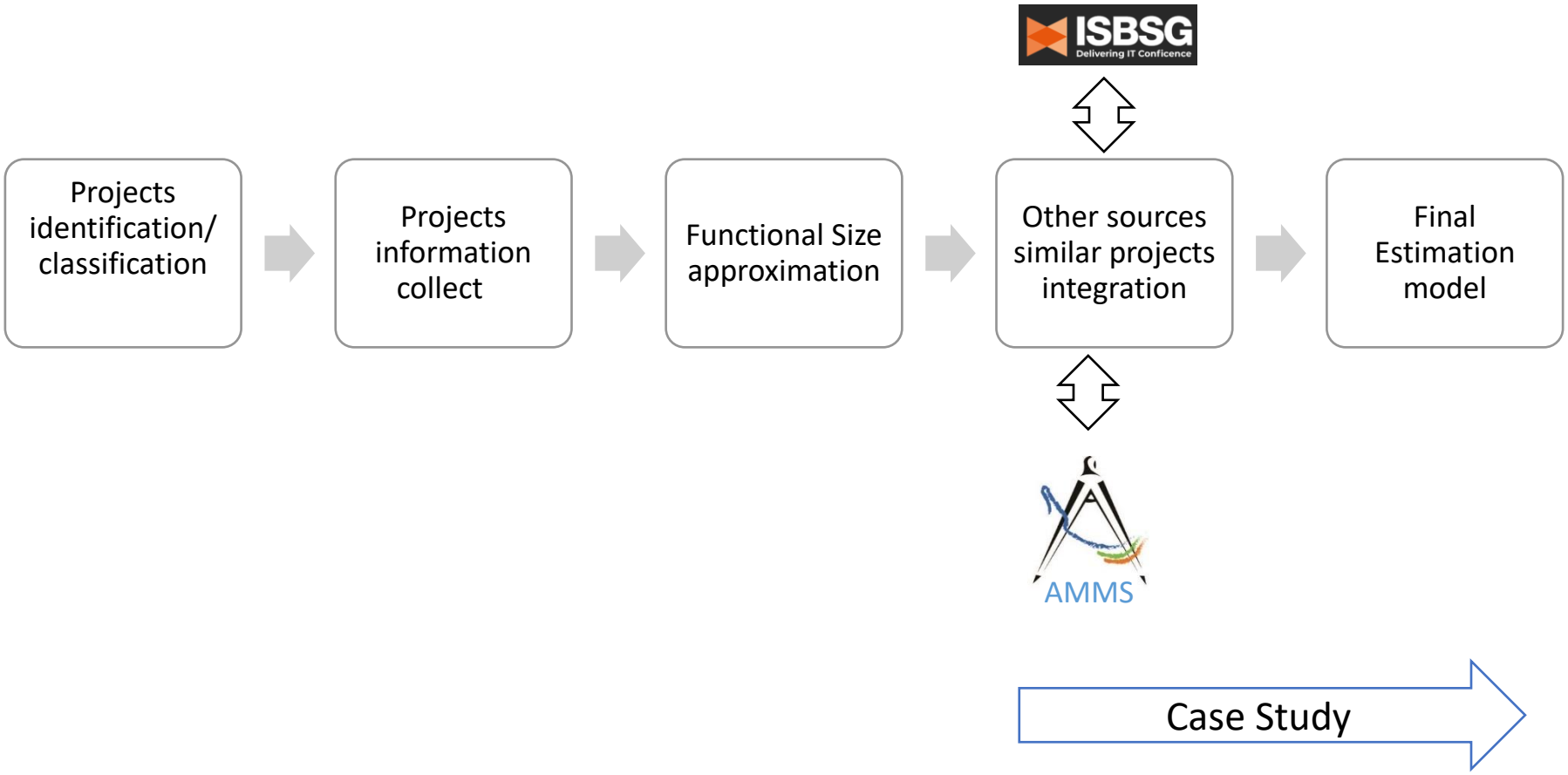
- Carbonera et al. [7] analyze the number of data points in the datasets and classify in **high quality** (more than 15 points) **medium quality** (10 to 15 data points and) or **low quality** (less than 10 data points), where it is possible to observe the **lack of datasets with a high number of data points, a main statistical principle**.

- Carbonera et al. find that **students or researchers are the most common participants** in the primary studies about effort estimation (91.67%). However, the presence of professionals is fundamental to produce realistic findings.

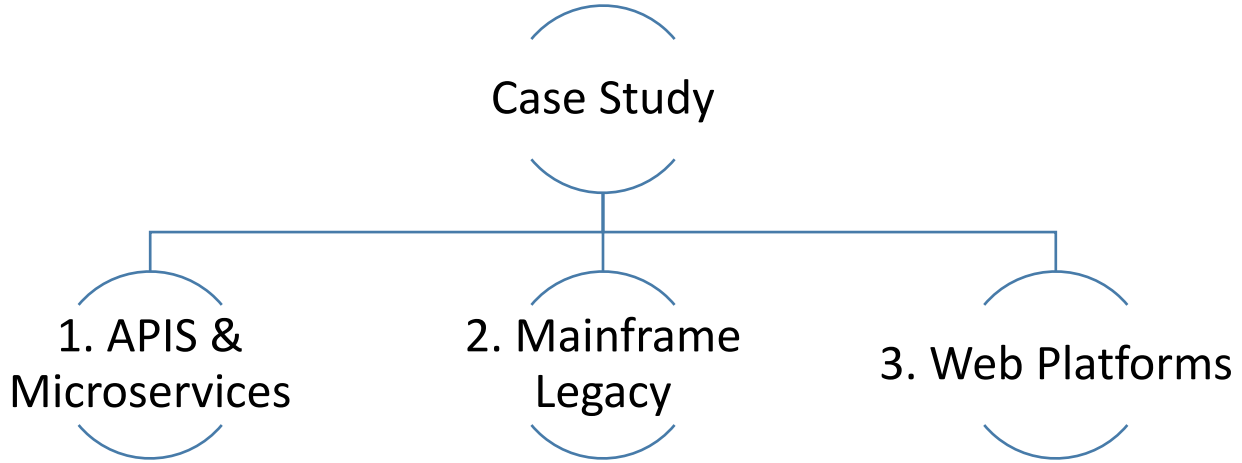
Problems in the Industry

- **Historical databases**
  - The absence of datasets using a FSMM
  - Effort/Cost required to generate a dataset
- **Scenarios**
  - Few data of projects
  - No data of projects

Steps for the Case Study



Projects in the Case Study



CLIENT	✓	✓	✗
ISBSG	✓	✓	✓
AMMS	✓	✗	✓

Case Study 1. APIS & Microservices

Data collect and FS Approximation

SOURCE	Sample Size	%
CLIENT	8	14.04%
ISBSG	15	26.31%
IMDS	34	59.65%
TOTAL	57	100.00%

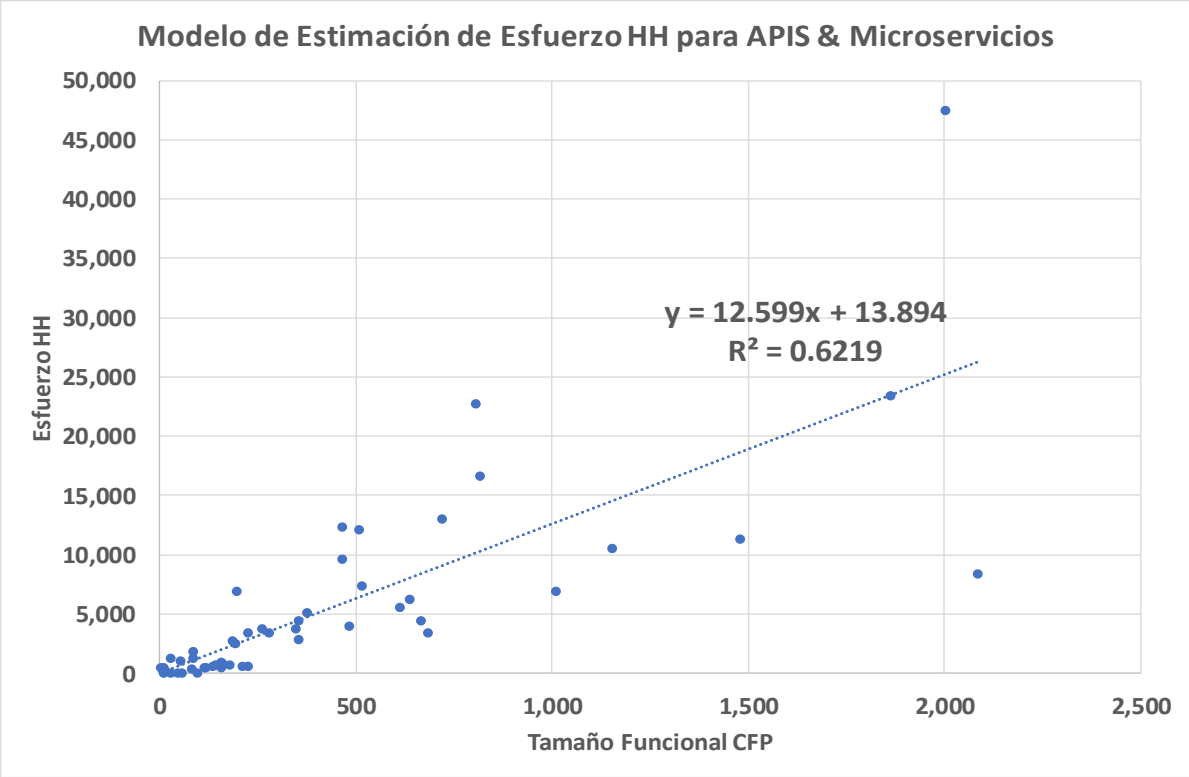
SOURCE	COSMIC Functional Size (CFP)	%
CLIENT	2,418.7	11.01%
ISBSG	3,873.0	17.63%
IMDS	15,674.6	71.36%
TOTAL	21,966.4	100%

PDR Analisys

Productivity	PDR
CFP/WH	WH/CFP
Productivity represents how many CFPs are implemented per work-hour	The PDR represents how many WH are required per CFP

SOURCE	Mín	P10	P25	Median	P75	P90	Máx	Media	DesvEst
CLIENT	12.3	----	12.5	14.2	20.9	----	35.5	17.7	8.0
ISBSG	0.2	0.3	2.3	3.9	4.7	12.9	23.7	4.6	5.6
IMDS	1.4	4.5	8.0	11.7	20.5	33.1	143.2	18.1	24.1
TOTAL	0.2	2.2	4.6	10.1	18.2	26.8	143.2	14.5	19.9

All the databases (57 projects)

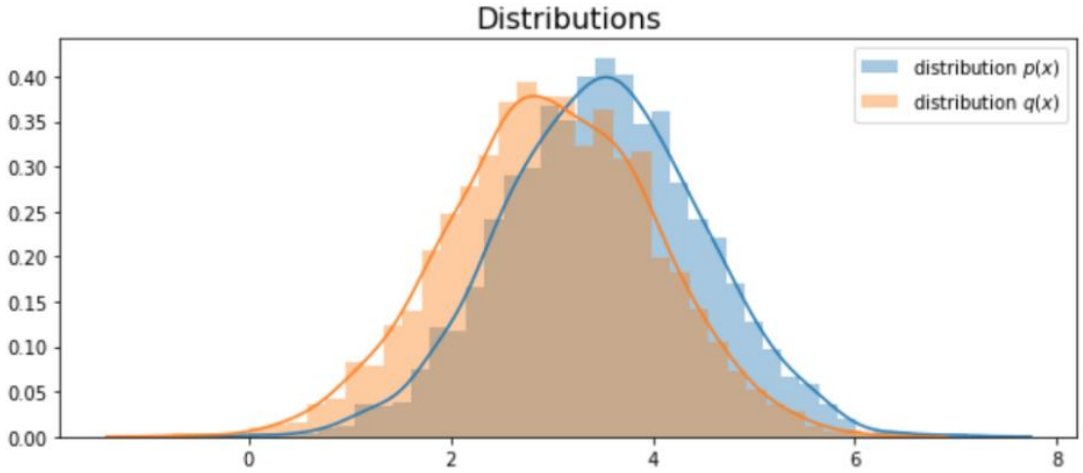


Question?

Could be integrated the three databases considering statistical foundations to get a high number of datapoints?

The integration make sense and it is valid?

Parametric test validation



- The **Kruskal-Wallis test**, also known as the H test, is the **non-parametric alternative to the one-way ANOVA test** for unpaired data.
- It is considered **an extension of the Mann-Whitney test for more than two groups**. It is therefore a test that uses ranges to contrast the hypothesis that k samples have been obtained from the same population.
- The Kruskal-Wallis test **contrasts whether the different samples are equidistributed and therefore belong to the same distribution** (**population**). In a simple way, the Kruskal-Wallis test compares the medians.

Parametric test - results

The test was developed considering the Product Delivery Rate (PDR), using SPSS ver 25 in Spanish.

Hypothesis:

- H0: Med1 = Med2 =… = Medk
- H1: Medi ≠ Medj for at least one pair (i, j)

N	**57**
Grados de libertad (número de agrupaciones -1)	2
Sig. Asintótica (p-value)	0.000

We can say that, since the **p-value** (Sig. Asymptot.) Is **lower than 0.05**, then **the null hypothesis (H0) is rejected** and it is concluded that with a **significance level of 5%,** there is a significant difference in at least one of the PDR distributions of the databases

Parametric test - results

To determine **which databases have different distributions**, it is necessary to perform an analysis using **the Kruskal-Wallis test in pairs**, adjusting the resulting p-value considering the number of tests, **this correction is known as the Bonferroni correction** for various tests.

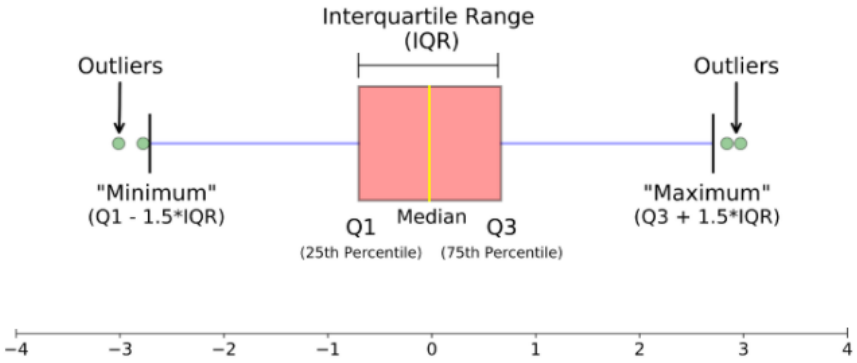
Pareja	Sig. Asintótica (p-value)	Sig. Asintótica (p-value) Ajustada [33]
ISBSG – IMDS	0.000	0.000
ISBSG – CLIENT	0.000	0.000
IMDS – CLIENT	**0.292**	**0.876**

We observe that **the IMDS - CLIENT pair** is the only one that has an adjusted p-value (0.876)> 0.05; from which we conclude that the **IMDS and CLIENT databases have the same distribution and could be integrated**.

Considering Outliers

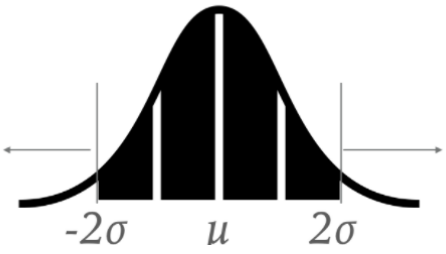
Inter quartile distance:

The method most taught academically for its simplicity and results is the **Tukey test**, which takes as a reference **the difference between the first quartile Q1 and the third quartile Q3**, or **the interquartile range IQR (Q3-Q1)**.

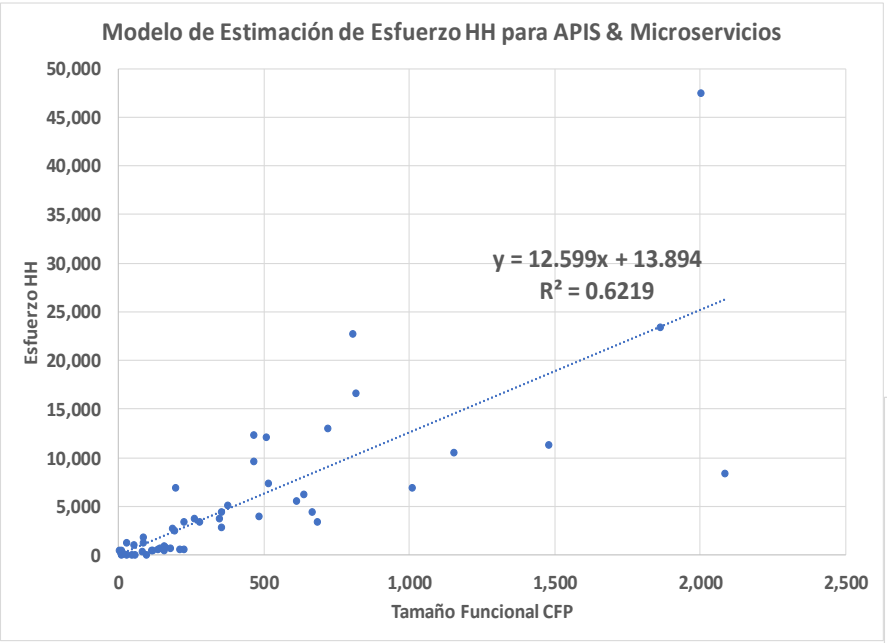


Using the sample standard deviation:

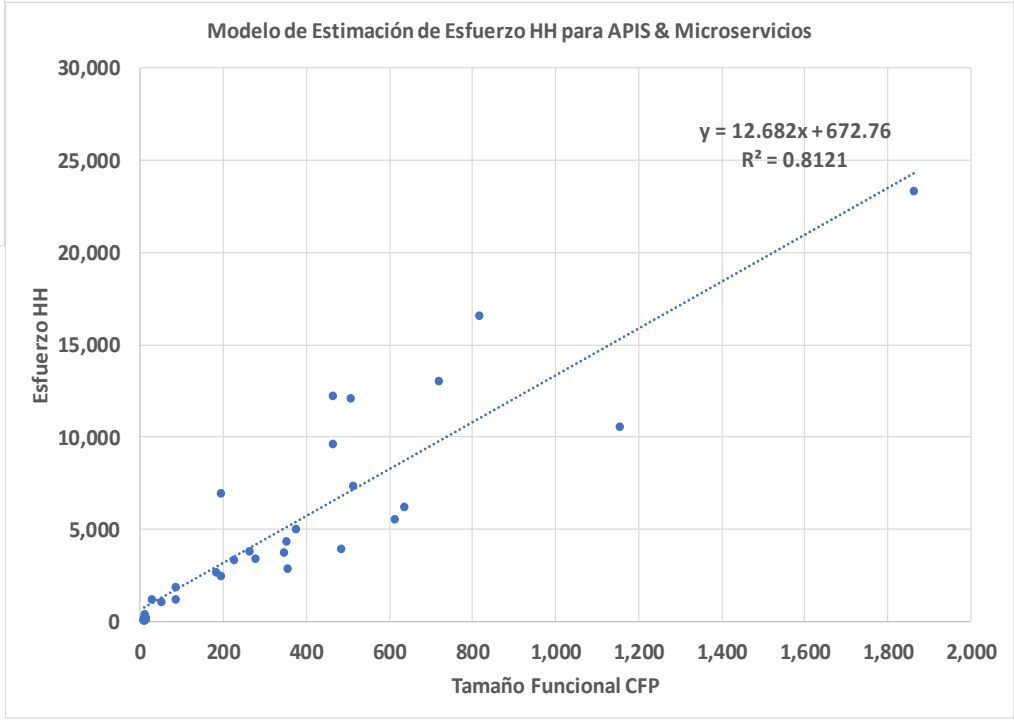
Another method used is to take the sample mean as a reference and all those points that are outside the interval of two standard deviations around the mean can be considered atypical; This method allows defining if we want to use 1, 2 or 3 standard deviations to be more or less flexible.



Final Estimation Model



CLIENT	**8**
IMDS	34
	42
Outliers	9
Total	33



Case Study 2.
Mainframe Legacy

Data collect and FS Approximation

SOURCE	Sample Size	%
CLIENT	10	22.2%
ISBSG	35	77.8%
IMDS	0	0.0%
TOTAL	45	100.0%

SOURCE	COSMIC Functional Size (CFP)	%
CLIENT	914.0	8.98%
ISBSG	9,267.0	91.02%
TOTAL	10,181.0	100%

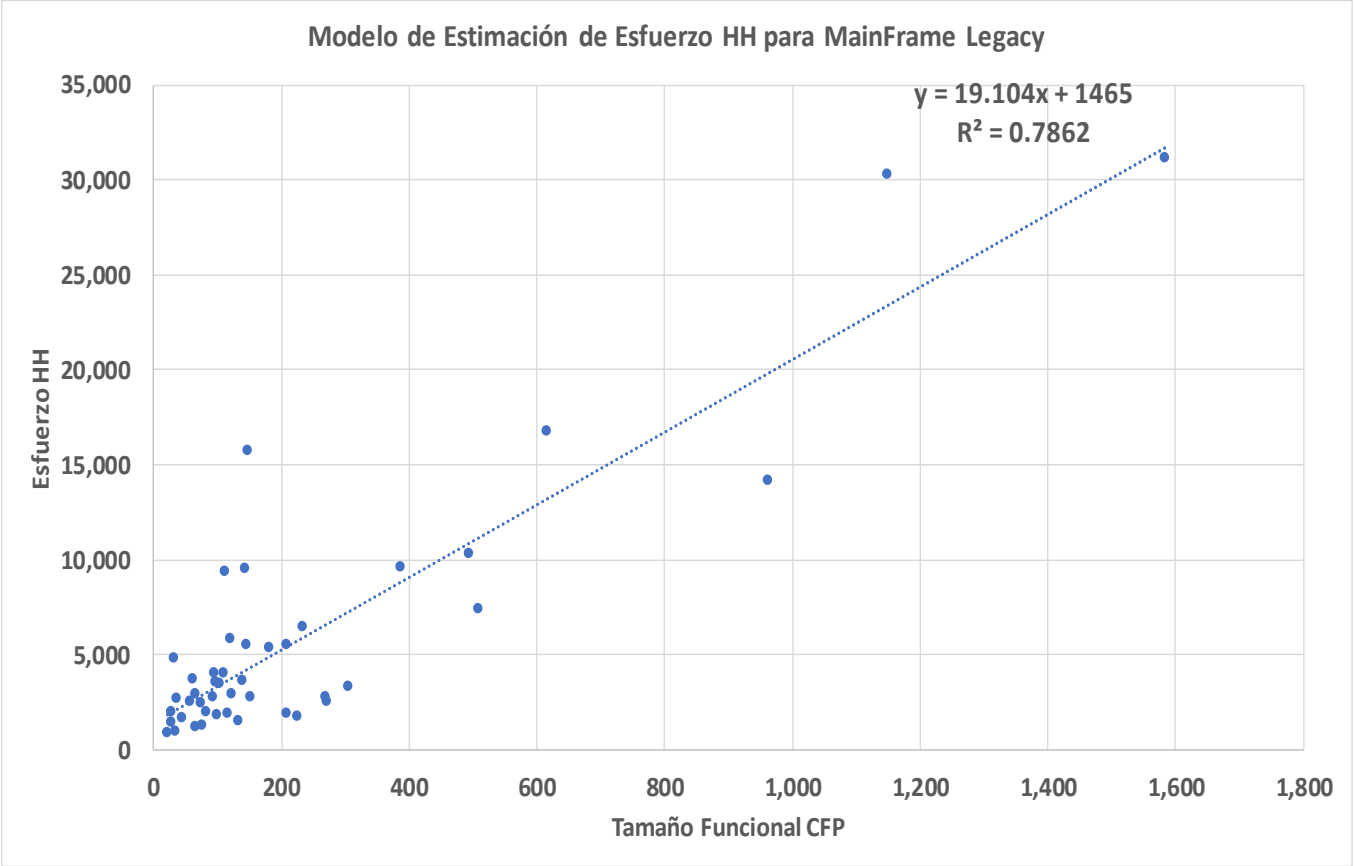
PDR Analisys

Productivity	PDR
CFP/WH	WH/CFP
Productivity represents how many CFPs are implemented per work-hour	The PDR represents how many WH are required per CFP

SOURCE	Mín	P10	P25	Median	P75	P90	Máx	Media	DesvEst
CLIENT	10.5	11.3	19.4	38.4	53.3	145.3	152.9	47.5	41.4
ISBSG	8.0	10.6	18.6	26.8	43.2	70.7	107.4	33.8	23.0
TOTAL	8.0	10.9	18.9	28.2	44.5	76.7	152.9	36.9	28.2

Facultad de Ciencias

All the databases (45 projects)



Question?

Could be integrated the three databases considering statistical foundations to get a high number of datapoints?

The integration make sense and it is valid?

Parametric test - results

The test was developed considering the Product Delivery Rate (PDR), using SPSS ver 25 in Spanish.

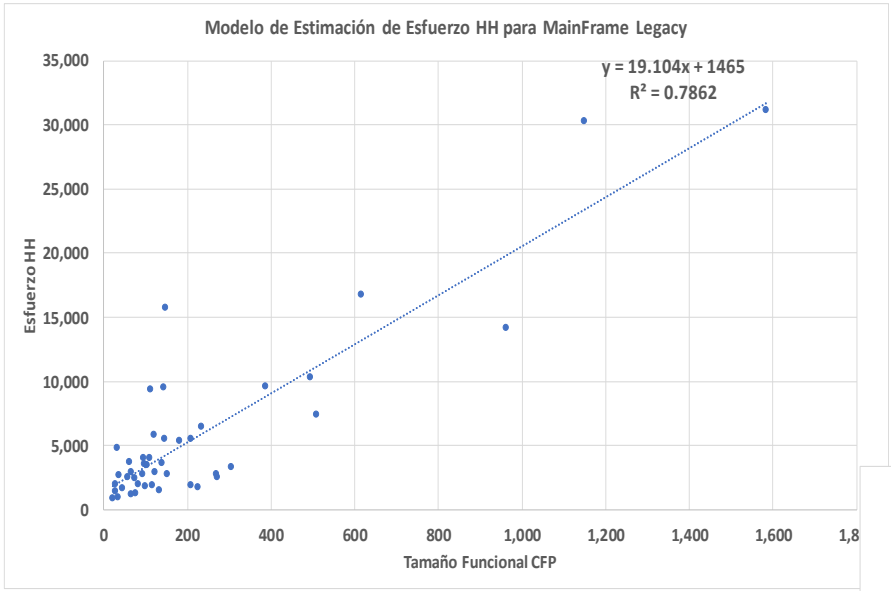
Hypothesis:

- H0: Med1 = Med2 =… = Medk
- H1: Medi ≠ Medj for at least one pair (i, j)

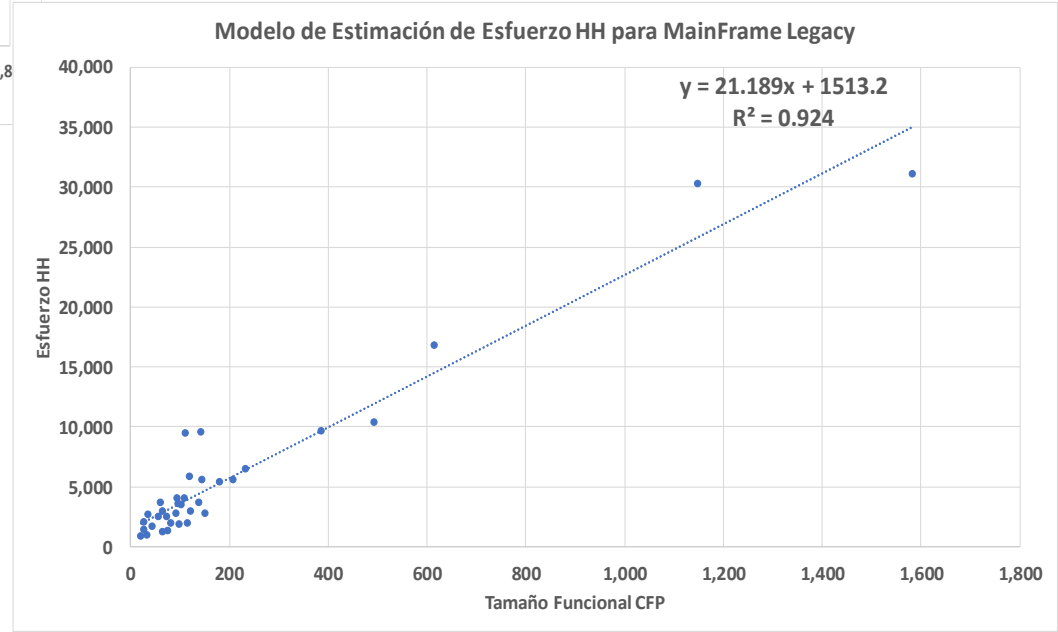
N	**45**
Grados de libertad (número de agrupaciones -1)	1
Sig. Asintótica (p-value)	0.275

We can say that, since the **p-value** (Sig. Asymptot.) Is **higher than 0.05**, then **the null hypothesis (H0) is accepted** and it is concluded that there is NO significant difference in the distributions of the CLIENT and ISBSG databases

Final Estimation Model



CLIENT	10
ISBSG	35
	45
Outliers	10
Total	35



Case Study 3.
Web Platforms

Data collect and FS Approximation

SOURCE	Sample Size	%
CLIENT	0	0.0%
ISBSG	20	44.4%
IMDS	25	55.6%
TOTAL	45	100.0%

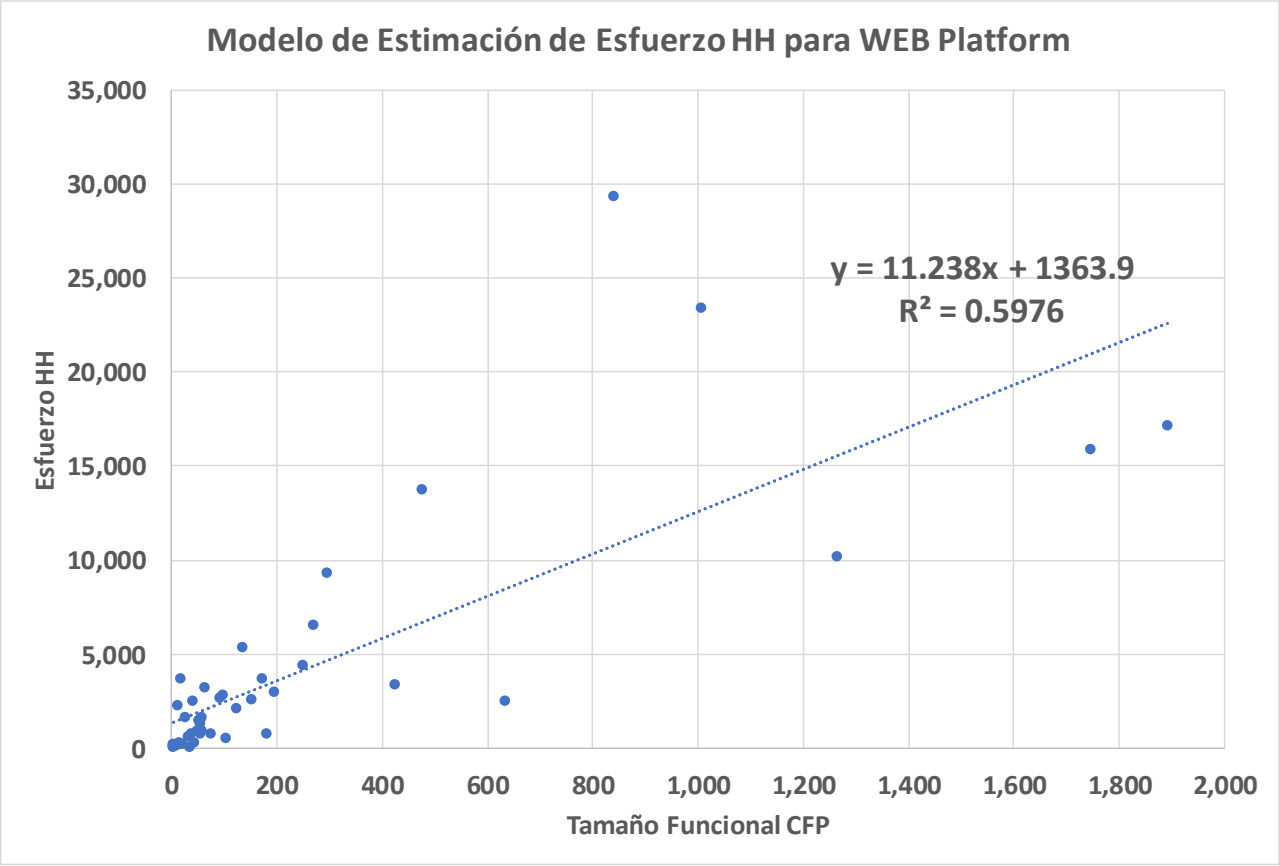
SOURCE	COSMIC Functional Size (CFP)	%
CLIENT	0.0	0.00%
ISBSG	3,721.0	33.31%
IMDS	7,449.1	66.69%
TOTAL	**11,170.1**	**100%**

PDR Analisys

Productivity	PDR
CFP/WH	WH/CFP
Productivity represents how many CFPs are implemented per work-hour	The PDR represents how many WH are required per CFP

SOURCE	Mín	P10	P25	Median	P75	P90	Máx	Media	DesvEst
ISBSG	4.0	13.9	18.3	24.8	48.5	217.2	257.8	50.4	68.9
IMDS	2.8	5.1	8.6	19.2	29.1	46.8	131.2	24.2	25.9
TOTAL	2.8	6.3	14.5	21.6	30.9	66.4	257.8	35.8	50.9

All the databases (45 projects)



Question?

Could be integrated the three databases considering statistical foundations to get a high number of datapoints?

The integration make sense and it is valid?

Parametric test - results

The test was developed considering the Product Delivery Rate (PDR), using SPSS ver 25 in Spanish.

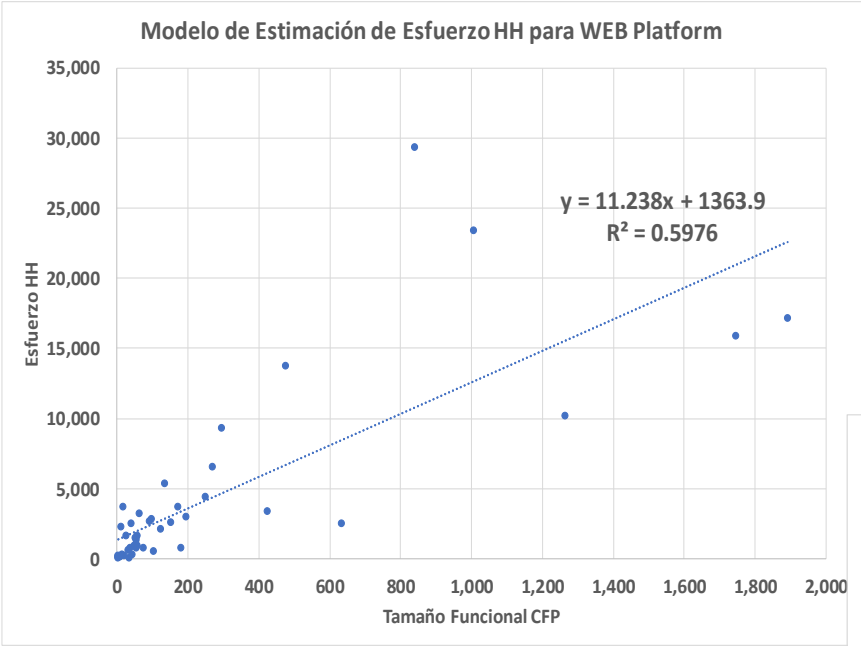
Hypothesis:

- H0: Med1 = Med2 =… = Medk
- H1: Medi ≠ Medj for at least one pair (i, j)

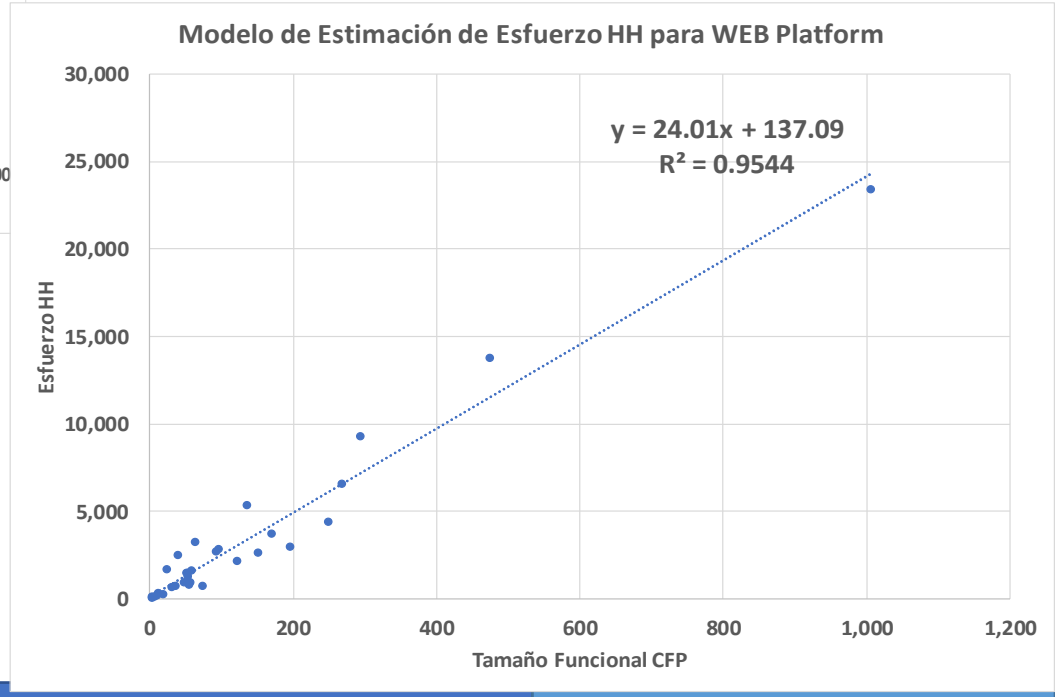
N	**45**
Grados de libertad (número de agrupaciones -1)	1
Sig. Asintótica (p-value)	0.054

We can say that, since the **p-value** (Sig. Asymptot.) Is **higher than 0.05**, then **the null hypothesis (H0) is accepted** and it is concluded that there is NO significant difference in the distributions of the IMDS and ISBSG databases

Final Estimation Model



IMDS	**25**
ISBSG	20
	45
Outliers	13
Total	32



Facultad de Ciencias

Conclusion

The use of the proposed procedure has made it possible to improve the estimation models in the Mexican industry from the integration of different databases, considering statistical foundations to validate the integration of different data sources.

ISBSG
IT Confidence 2021

Questions?

Dr. Francisco Valdés-Souto
Associate Professor
Department of Mathematics,
Science Faculty,
National Autonomous University of Mexico (UNAM)
fvaldes@Ciencias.unam.mx

COSMIC President

COSMIC