ANÁLISIS DE RIESGO DE IRREGULARIDAD EN CONTRATACIÓN PÚBLICA COLOMBIANA

Sebastián Tutistar Valencia - Miguel Angel Rincón

Avance #1

Pregunta de Investigación

Agencia Nacional
de Contratación Pública
Colombia Compra Eficiente

¿Cuál es el nivel poblacional estimado de riesgo de irregularidad en los procesos de contratación pública adjudicados en Colombia durante el periodo 2020–2024, diferenciado por departamento y modalidad de contratación, a partir de un diseño de muestreo probabilístico aplicado al marco muestral del SECOP?

Se busca estimar estos parámetros mediante un diseño muestral probabilístico:

- ¿Qué departamentos presentan mayores indicadores de riesgo en la contratación pública?
- ¿Existen diferencias en el nivel de riesgo según la modalidad de contratación utilizada?

Objetivos del Estudio

Estimar el índice promedio de riesgo de irregularidades en la contratación pública por departamento y modalidad de contratación en Colombia (2020-2024), mediante un diseño de muestreo estratificado con estimación por dominios, a partir de los datos de SECOP.

- Comparar diseños muestrales (MAS vs Estratificado) y justificar la elección
- Calcular el tamaño de muestra óptimo considerando precisión y costos
- Estimar el IRI a nivel nacional y por dominios (Departamento × Modalidad)

Población y Marco Muestral

Población objetivo: Todos los procesos de contratación pública adjudicados registrados en SECOP durante 2020-2024. Población Total: N = 583.198 contratos.

Marco Muestral:

- Fuente: Base de datos SECOP (datos.gov.co)
- Unidad de muestreo: Proceso contractual individual
- Estratos: 33 (32 departamentos + Bogotá D.C.)
- Dominios de estudio: Departamento × Modalidad
- Período: 1 enero 2020 - 31 diciembre 2024

Justificación del Muestreo:

Dada la magnitud del marco, un censo implicaría costos computacionales, tiempo de procesamiento y complejidad innecesarios. El muestreo estratificado permite obtener estimaciones precisas con recursos limitados.

Diseño de Muestreo Propuesto M A S

Estratificación:

- Selección directa de unidades (contratos) de todo el marco muestral sin estratificación previa.

Propiedades del MAS:

- Sencillo de aplicar, replicar y documentar.
- Insesgado para estimar medias y proporciones poblacionales.
- Varianza relativamente alta cuando la población es heterogénea (no aprovecha diferencias entre departamentos o modalidades).
- Menor eficiencia frente a diseños estratificados si existen grupos naturalmente distintos.

Marco Muestral:

- Población total: N contratos.
- Cada contrato tiene igual probabilidad de selección.
- Requisito clave: listado completo, depurado y sin duplicados.

Diseño: MAS (Muestreo Aleatorio Simple) en cada departamento.

Sin reemplazo.

- Cada contrato tiene $\pi = n_h/N_h$ de ser seleccionado.
- Permite estimación insesgada.

Diseño de Muestreo Propuesto Muestreo Estratifcado

Estratificación:

- Variable de estratificación: Departamento
- H = 33 estratos
- Justificación: Reduce varianza al agrupar unidades homogéneas geográficamente
- Los departamentos varían en volumen de contratación (N_1, N_2, ..., N_{33})

Muestreo dentro de estratos:

- Diseño: MAS (Muestreo Aleatorio Simple) en cada departamento.

Sin reemplazo.

- Cada contrato tiene $\pi = n_h/N_h$ de ser seleccionado.
- Permite estimación insesgada.

Estimación por Dominios:

Los dominios de estudio son las combinaciones de depto x modalidad = 33 x 13 = 429 dominios.
Después de seleccionar la muestra estratificada, se analizará el IRI promedio para cada dominio (ej: "Licitación Pública en Antioquia", "Contratación Directa en Valle", etc.)

$$D_{posibles} = 33 \times 13 = 429$$
$$D_{observados} = \sum I(n_{dm} > 0) = 323$$
$$D_{estimables} = \sum I(n_{dm} \geq 5) = 296$$

Tamaño de Muestra

SECOP II - Procesos de Contratación

Estadísticas Nacionales

Registro de los procesos de compra, sean o no adjudicados, hechos en la plataforma SECOP II desde su lanzamiento

Última Actualización
23 de noviembre de 2025

Datos suministrados por
Colombia Compra Eficiente

Información sobre este conjunto de datos

Actualizado
23 de noviembre de 2025

Última actualización de los datos
23 de noviembre de 2025

Última actualización de metadatos
27 de octubre de 2025

Fecha de creación
30 de septiembre de 2019

Vistas
1,37M

Descargas
109K

Suministró los datos
Colombia Compra Eficiente

Propietario de conjunto de datos
Datos Abiertos CCE

Información de la Entidad

Departamento	Bogotá D.C.
Municipio	Bogotá D.C.
DIVIPOLA Municipio	11001
Nombre de la Entidad	Agencia Nacional De Contratación Pública - Colombia Compra Eficiente, Bogotá D.C.
Orden	Nacional
Área o dependencia	Subdirección de IDT
Sector	Planeación

Información de Datos

Cobertura Geográfica	Nacional
Idioma	Español
Frecuencia de Actualización	Diaria
Fecha Emisión (aaaa-mm-dd)	2019-10-01

Elementos adjuntos

14_Diccionario_de_Datos- SECOP II - Procesos de Contratación.docx

¿Qué hay en este conjunto de datos?

Filas	Columnas	Cada fila es un
8,22M	**59**	**Proceso**

Columnas (59)

Nombre de la columna	Descripción

8 Millones

2020-2024

Contratos

contratos adjudicados

N = 583,198

Variables - Clasificación por Rol

Categoría (Rol)	Variables	Cantidad
Variable Principal	IRI	1
Diseño Muestral	departamento_entidad, modalidad_de_contratacion	2
Construcción del IRI	proveedores_unicos_con_respuestas, proveedores_con_invitacion_directa,	5
Identificación	id_del_proceso, entidad, nombre_del_proveedor_adjudicado,	4
Temporales	fecha_de_publicacion_del_proceso, fecha_adjudicacion	2
No utilizadas en el análisis	adjudicado, estado_del_procedimiento, estado_resumen, ciudad_entidad,	5
—	**Total variables en dataset original**	**59**
—	**Variables seleccionadas para el estudio**	**17**
—	**Variables realmente usadas en la estimación (IRI + diseño + construcción)**	**8**

Total de variables en el dataset original: 59

Variables seleccionadas: 17

Variables realmente usadas en estimación: 8 (IRI + 2 diseño + 5 construcción)

Variable Principal

IRI - Índice de Riesgo de Irregularidad

- Tipo: Cuantitativa continua
- Rango: [0, 1]
- Construcción: Variable compuesta derivada de 5 banderas binarias
- Fórmula: IRI = 0.25×flag1 + 0.25×flag2 + 0.15×flag3 + 0.15×flag4 + 0.20×flag5

Es la variable que se ESTIMA a nivel nacional y por dominios.

#	Variable	Tipo	Rol	Uso en IRI	Valores Únicos
1	modalidad_de_contratacion	Nominal	Dominio	flag1	13
2	proveedores_unicos_con_respuestas	Discreta	Auxiliar	flag2	261
3	proveedores_con_invitacion_directa	Discreta	Auxiliar	flag3	39
4	duracion	Continua	Auxiliar	flag4	
5	valor_total_adjudicacion	Continua	Auxiliar	flag5	

Sesgos Potenciales y Limitaciones

- Sesgo de clasificación: La definición de "riesgo" se basa en indicadores indirectos, no en comprobación judicial de corrupción.

- Sesgo de cobertura: No todos los procesos de contratación pueden estar registrados o completos, especialmente en municipios con menor capacidad institucional, limitaciones técnicas o baja conectividad, lo que puede subrepresentar ciertos territorios o modalidades.

- Sesgo de agregación por dominio: Al trabajar con dominios Departamento × Modalidad, algunos dominios poco frecuentes se excluyen por tamaño muestral insuficiente (n>5), limitando inferencia en combinaciones raras.

Sesgos Potenciales y Limitaciones

Limitaciones logísticas:

- Calidad de datos: Inconsistencias, campos vacíos o errores en los registros disponibles en datos.gov.co.
- Volumen de datos: Procesamiento de millones de registros requiere capacidad computacional significativa.
- Disponibilidad del portal: Dependencia de la disponibilidad y estabilidad del portal datos.gov.co para la consulta de información.

Cronógrama

Fase	Actividad	Descripción detallada	Duración	Período
Fase 1	Consulta de datos desde datos.gov.co	Identificación de los conjuntos de datos relevantes en el portal datos.gov.co, consulta y extracción de información de SECOP para el período 2020-2024, verificación de integridad de la información obtenida y documentación de metadatos (fecha de consulta, versión del dataset, URL de origen).	2 días	Nov 24-25
Fase 2	Limpieza y preparación de datos	Exploración inicial de la estructura de los datos, identificación y tratamiento de valores faltantes, detección y corrección de inconsistencias (duplicados, errores de formato, valores atípicos), estandarización de variables categóricas (departamentos, modalidades, sectores), conversión de tipos de datos y creación de base de datos consolidada.	3 días	Nov 26-28
Fase 3	Construcción de variables e índice de riesgo	Ingeniería de variables derivadas (proporciones, indicadores temporales, categorías de cuantía), identificación y construcción de red flags (banderas de alerta), cálculo del Índice de Riesgo de Irregularidad (IRI) mediante ponderación de indicadores, validación técnica del índice con pruebas estadísticas y ajuste de metodología según hallazgos preliminares.	3 días	Nov 29 - Dic 1

Cronógrama

Fase 4	Análisis exploratorio y descriptivo	Estadísticas descriptivas por departamento, modalidad y año, identificación de patrones temporales y geográficos, creación de visualizaciones exploratorias (mapas de calor, series temporales, distribuciones), análisis de correlaciones entre variables, detección de datos atípicos y casos especiales que requieran atención.	3 días	Dic 2-4
Fase 5	Tècnicas de Muestreo y Estimaciòn	Aplicación de los estimadores propios del diseño (MAS y estratificado), cálculo de medias, proporciones y totales, estimación de varianzas, construcción de intervalos de confianza, factores de expansión, ajustes por no respuesta y calibración. Validación de la precisión y consistencia de las estimaciones entre dominios.	4 días	Dic 5-8
Fase 6	Validación y análisis de sensibilidad	Análisis de robustez del índice de riesgo bajo diferentes especificaciones, validación cruzada de hallazgos con fuentes secundarias (informes de organismos de control), ajustes finales a metodología y conclusiones.	2 días	Dic 9-10

Cronógrama

Fase 7	Elaboración de informe y visualizaciones finales	Redacción del informe final con estructura académica completa, creación de visualizaciones de alta calidad para presentación (mapas interactivos, dashboards, gráficos), elaboración de resumen ejecutivo y conclusiones, preparación de material de presentación (slides), revisión final y correcciones de forma y fondo.	2 días	Dic 11-12
Presentación	**Presentación final**	**Entrega y sustentación del trabajo final de investigación**	**1 día**	**13 dic 2025**

Avance #2

Comparación de diseños muestrales (MAS y Estratificado)

Muestreo Aleatorio Simple

- Selección aleatoria directa
- Sin estratificación geográfica
- Cobertura: 33/33 departamentos
- Dominios estimables (n≥5): 82
- DEFF: 1.0 (referencia)

Muestreo Estratificado

- Estratificación por departamento
- Afijación proporcional al tamaño
- Cobertura: 33/33 departamentos
- Dominios estimables (n≥5): 92
- DEFF: 0.85 (15% más eficiente)

Diseño elegido: Estratificado

Mejor precisión estadística (DEFF=0.85) y garantiza representatividad en todos los departamentos.
Reduce varianza entre estratos aprovechando la homogeneidad dentro de cada departamento.

Cálculo del Tamaño de Muestra

Párametros

- Error máximo (e): 5%
- Nivel de Confianza: 95% (z=1.96)
- Proporción esperada (p): 30%
- DEFF (efecto diseño): 0.85
- Tasa no respuesta: 10%
- Población (N): > 500.000

Tamaño inicial para proporción

$$n_0 = \frac{z^2\, p(1-p)}{(e \cdot p)^2}$$

Corrección por población finita, DEFF y no respuesta.

$$n = \frac{n_0 \cdot DEFF \cdot N}{n_0 \cdot DEFF + N - 1} \cdot \frac{1}{1 - NR}$$

Resultado Final

n = 3,369 contratos

Distribuidos proporcionalmente entre los 33 departamentos.

Justificación de los parámetros utilizados

Error Relativo del 5%: Se busca estimar el IRI con un margen de error relativo del 5%. Si el IRI verdadero es 16%, el error absoluto máximo será ±0.8 puntos porcentuales, proporcionando estimaciones precisas para identificar dominios de alto riesgo.

Confianza del 95%: Nivel estándar en investigación científica (z=1.96). Implica que en 95 de cada 100 muestras, el intervalo de confianza contendrá el verdadero valor del parámetro poblacional.

DEFF = 0.85: El diseño estratificado es 15% más eficiente que el MAS. Esto ocurre porque la estratificación por departamento reduce la variabilidad entre unidades dentro de cada estrato, mejorando la precisión sin aumentar el tamaño muestral.

Tasa de No Respuesta = 10%: Supuesto conservador basado en estudios con registros administrativos. El análisis demostró que el impacto real fue despreciable: Tasa de respuesta obtenida: 92.2%
Diferencia en IRI ajustado vs. no ajustado: 1.01% → Impacto DESPRECIABLE en las estimaciones

Asignación proporcional en muestreo estratificado

Cada departamento recibe una porción de la muestra proporcional a su tamaño en la población.

donde:

- nh: tamaño de muestra del estrato h
- n: tamaño total de la muestra
- Nh: tamaño del estrato h
- N: tamaño total de la población

$$n_h = n \times \frac{N_h}{N}$$

Ventajas:

- Representatividad geográfica.
- Cada departamento contribuye proporcionalmente.
- Facilita comparaciones entre dptos.

Resultados:

- 33 estratos (departamentos)
- 3,369 unidades totales
- Cobertura: 100% de deptos.
- 82 dominios con n≥5
- Estimación nacional: CV=1.89%

Modelado de Costos

$$CT_0 = C_0 + C_1 \times n$$

donde:
C_0: Costo Fijo (Diseño del estudio, infraestructura, software)
C_1: Costo Variable (Costo por unidad muestral procesada)
n: Tamaño Muestra (Número de contratos a analizar)

A diferencia de encuestas tradicionales, el uso de registros administrativos como SECOP elimina los costos de campo, que típicamente representan 60-80% del presupuesto en encuestas presenciales. Los costos se concentran en procesamiento y análisis de datos.

Estructura de Costos

Concepto	MAS	Estratificado
Costos fijos (C_0)		
Diseño del estudio	$150	$200
Extracción base de datos	$100	—
Extracción estratificada (34 dptos)	—	$150
Configuración software / ponderadores	$50	$100
Subtotal C_0	**$300**	**$450**

Costo variable por contrato (C_1)	MAS	Estratificado
Limpieza de datos	$0.10	$0.10
Cálculo IRI	$0.05	$0.05
Validación	$0.03	—
Validación + peso por estrato	—	$0.05
C_1 por contrato	$0.18	$0.20

Costo procesamiento muestra (n = 3,369)	MAS	Estratificado
$C_0 + (C_1 \times n)$	**$906**	**$1,123.80**
Análisis estadístico	$200	—
Informe resultados	$200	—
Análisis por estratos	—	$250
Informe desagregado	—	$300
Costo total	**$1,306**	**$1,674**
Costo por contrato	**$0.39**	**$0.50**

Análisis de Costo - Beneficio

Indicador	Resultado	Interpretación
Diferencia absoluta de costo	$368	Incremento total frente a MAS
Incremento porcentual	28.20%	Costo adicional moderado
Ganancia en precisión	15%	Reducción de varianza (DEFF = 0.85)
n efectivo ganado	+594 observaciones	Mayor potencia estadística
Costo por observación efectiva	$0.42	1,674 / 3,963 obs efectivas

Desglose de la Diferencia de $368

Componente	Incremento
Mayor C_0 (diseño más complejo)	$150
Mayor C_1 ($0.20 vs $0.18)	$67
Análisis desagregado por estratos	$150
Total diferencia	**$368**

El muestreo estratificado justifica su sobrecosto de $368 (28.2%) al proporcionar 15% más de precisión estadística(equivalente a 594 observaciones adicionales), garantizar representatividad en los 33 departamentos, y generar 82 dominios estimables para realizar por ejemplo, estudios focales en la contratación pública.

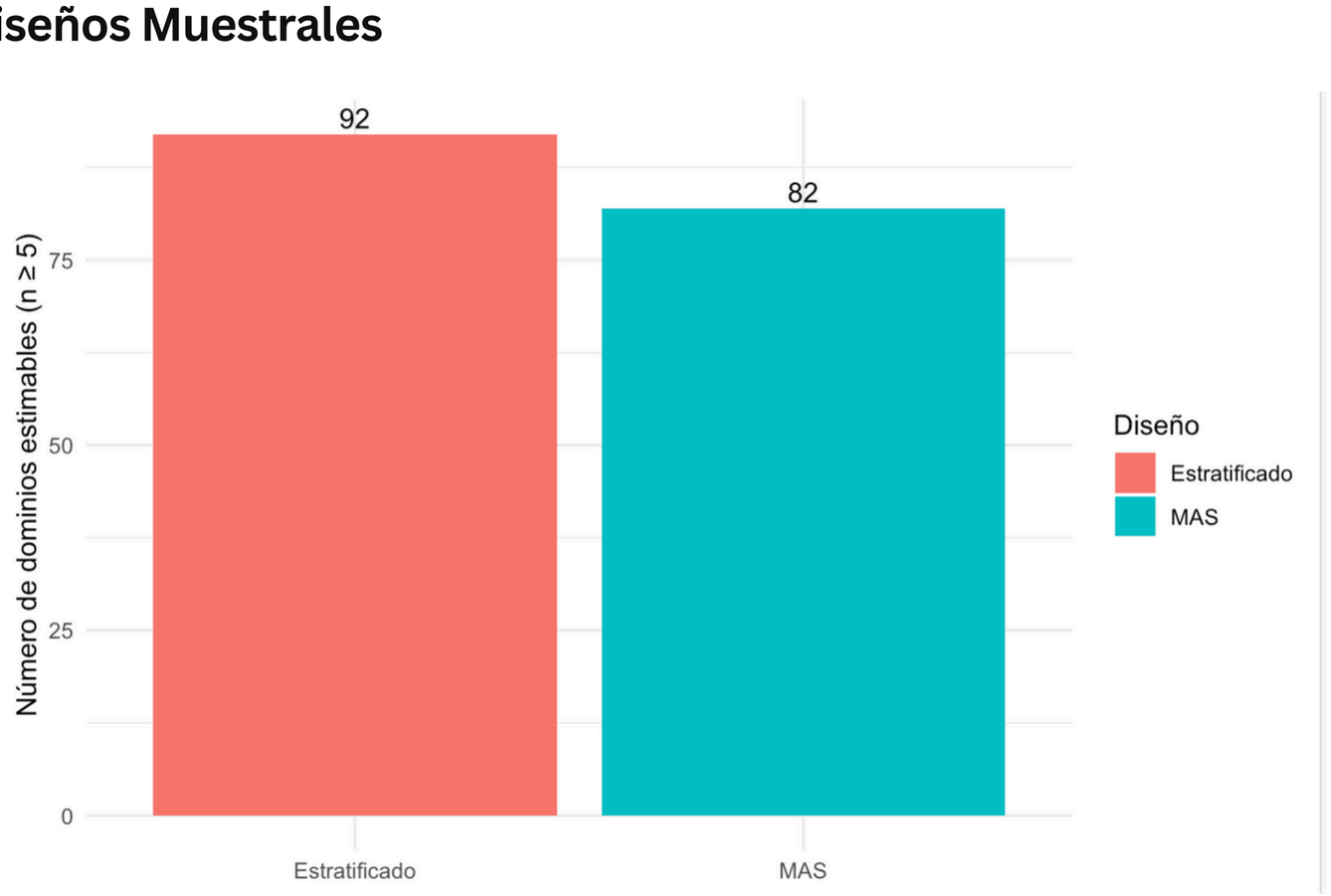
El costo marginal por observación es de solo $0.034(+8.8%), lo que representa una excelente relación costo-beneficio para los objetivos del estudio.

Avance #3

Graficos y tablas de resultados

Comparación de Diseños Muestrales

Diseño	Cobertura_Departamentos	Dominios_n_ge_5	DEFF
MAS	33	82	1.00
Estratificado	33	92	0.85

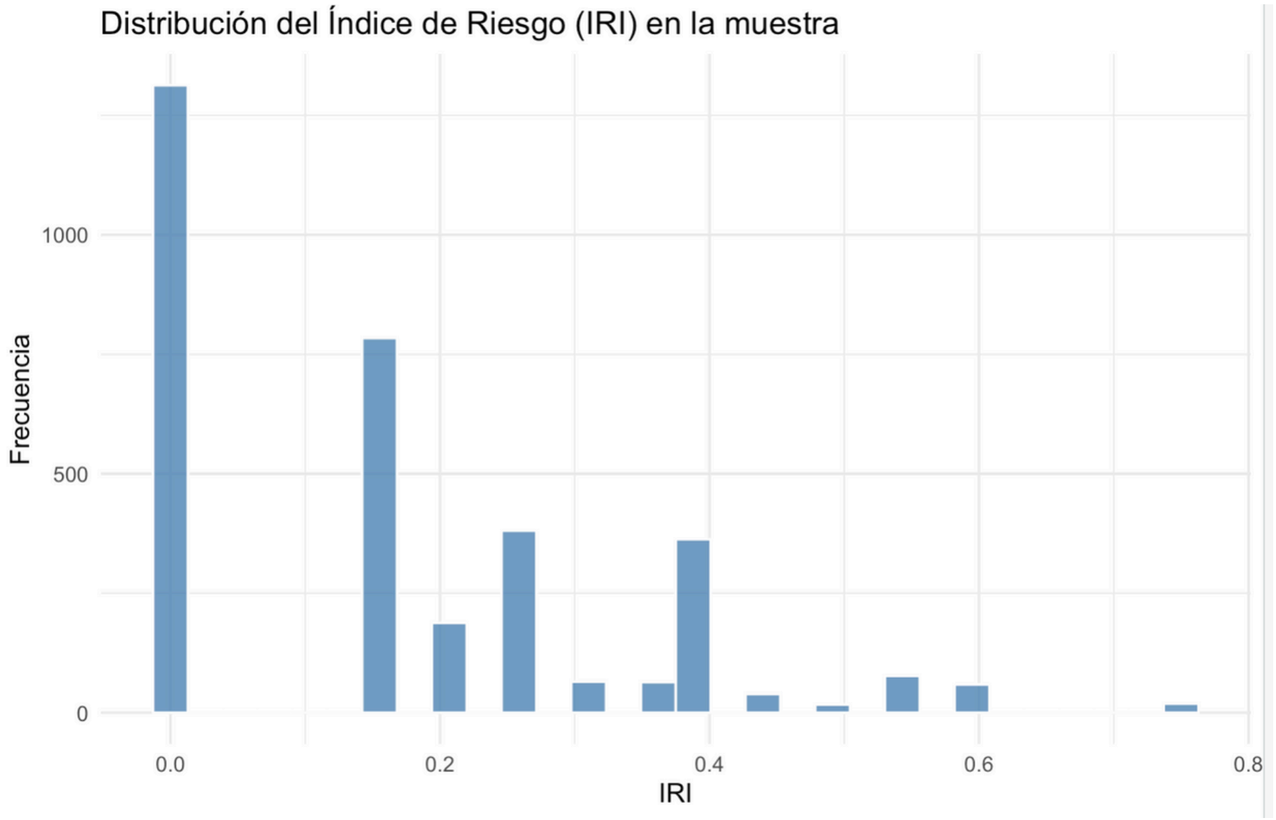


Graficos y tablas de resultados

Estimación Indice de riesgo

Estimación Nacional del IRI

Indicador	Valor
IRI Nacional	0.1636
Error Estándar	0.0031
Coeficiente de Variación (%)	1.8900
IC 95% (Límite Inferior)	0.1576
IC 95% (Límite Superior)	0.1697

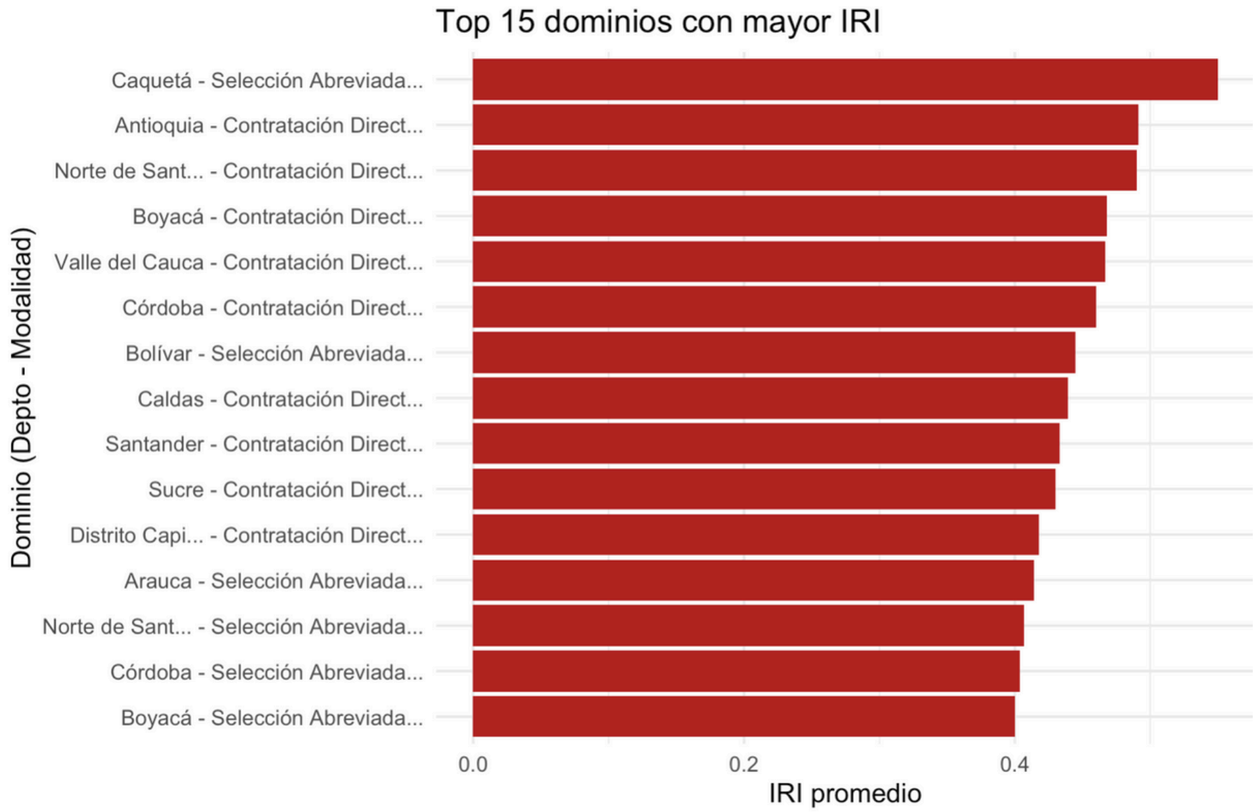


Graficos y tablas de resultados

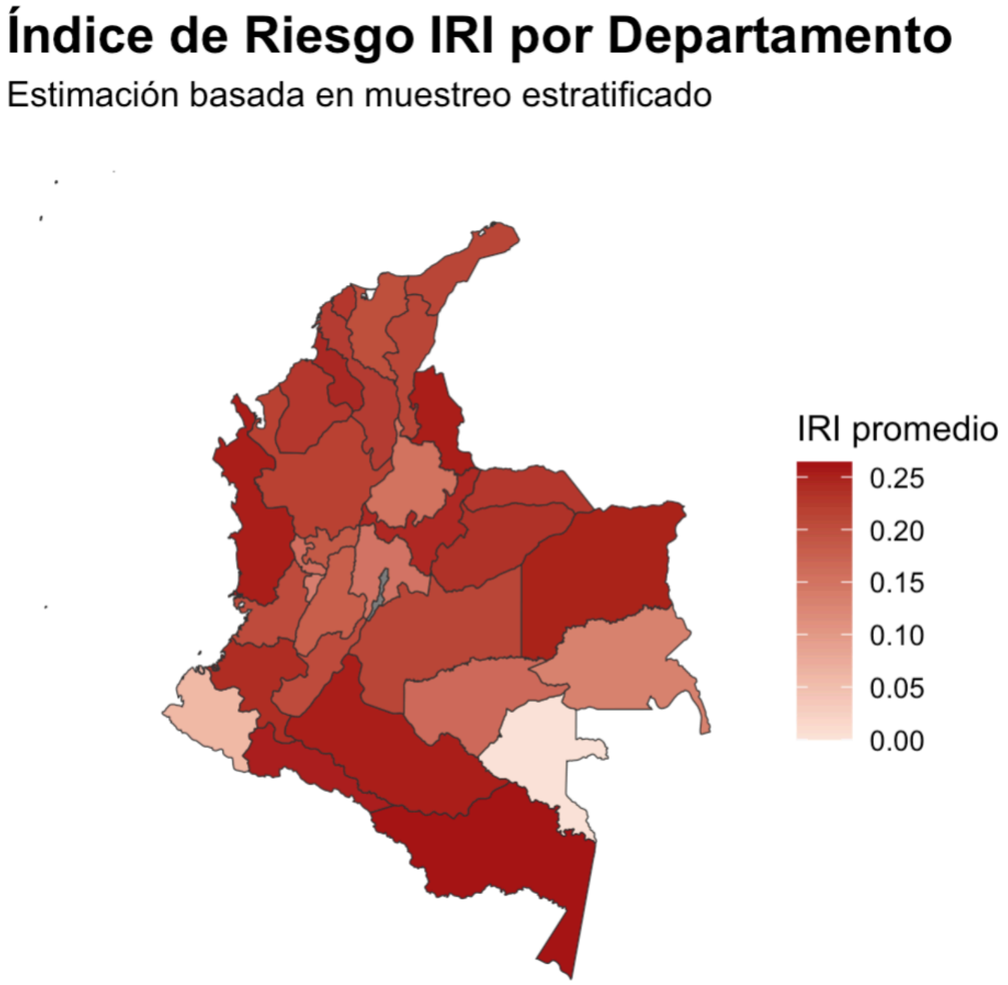
Estimación Indice de riesgo

Top 15 Dominios con Mayor Riesgo (IRI)

#	Depto	Modal	n	IRI	SD	Rango	CV%	Método	AmplIC	Prec
1	Caquetá	Selección Abreviada...	5	0.550	0.000	0.00	0.0	Constante	0.000	★★★★ Perfecta
2	Antioquia	Contratación Direct...	58	0.491	0.138	0.60	28.2	SD	0.071	★★ Buena
3	Norte de Sant...	Contratación Direct...	5	0.490	0.082	0.15	16.8	SD	0.144	★★ Buena
4	Boyacá	Contratación Direct...	11	0.468	0.096	0.20	20.4	SD	0.113	★★ Buena
5	Valle del Cauca	Contratación Direct...	18	0.467	0.128	0.60	27.5	SD	0.119	★★ Buena
6	Córdoba	Contratación Direct...	15	0.460	0.111	0.35	24.0	SD	0.112	★★ Buena
7	Bolívar	Selección Abreviada...	11	0.445	0.099	0.30	22.1	SD	0.117	★★ Buena
8	Caldas	Contratación Direct...	9	0.439	0.078	0.20	17.8	SD	0.102	★★ Buena
9	Santander	Contratación Direct...	15	0.433	0.070	0.20	16.1	SD	0.071	★★ Buena
10	Sucre	Contratación Direct...	5	0.430	0.067	0.15	15.6	SD	0.118	★★ Buena
11	Distrito Capi...	Contratación Direct...	98	0.418	0.161	0.60	38.5	SD	0.064	★ Aceptable
12	Arauca	Selección Abreviada...	7	0.414	0.189	0.40	45.5	SD	0.280	★ Aceptable
13	Norte de Sant...	Selección Abreviada...	7	0.407	0.195	0.60	47.8	SD	0.288	★ Aceptable
14	Córdoba	Selección Abreviada...	13	0.404	0.148	0.45	36.6	SD	0.161	★ Aceptable
15	Boyacá	Selección Abreviada...	5	0.400	0.000	0.00	0.0	Constante	0.000	★★★★ Perfecta

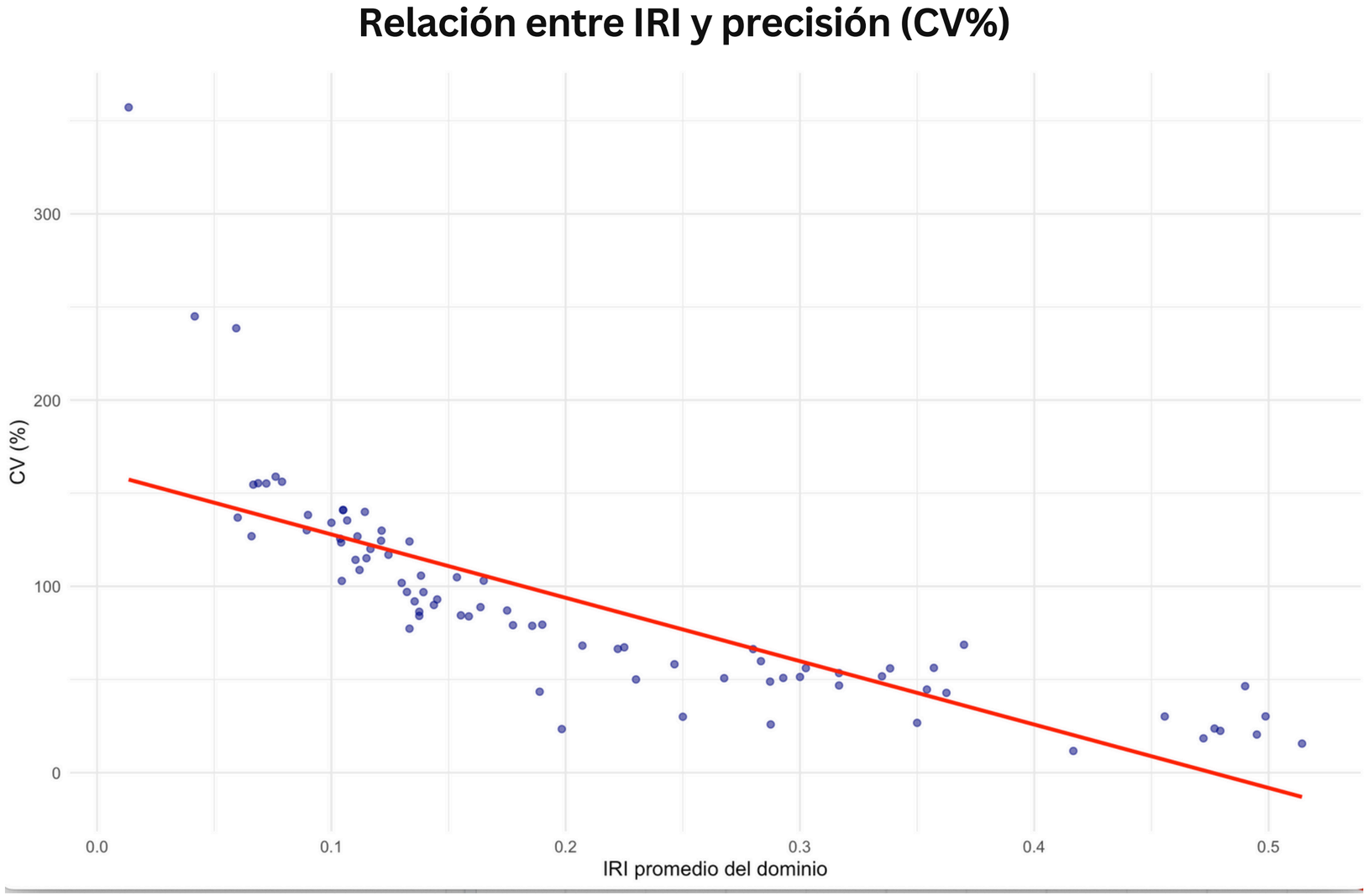


Graficos y tablas de resultados

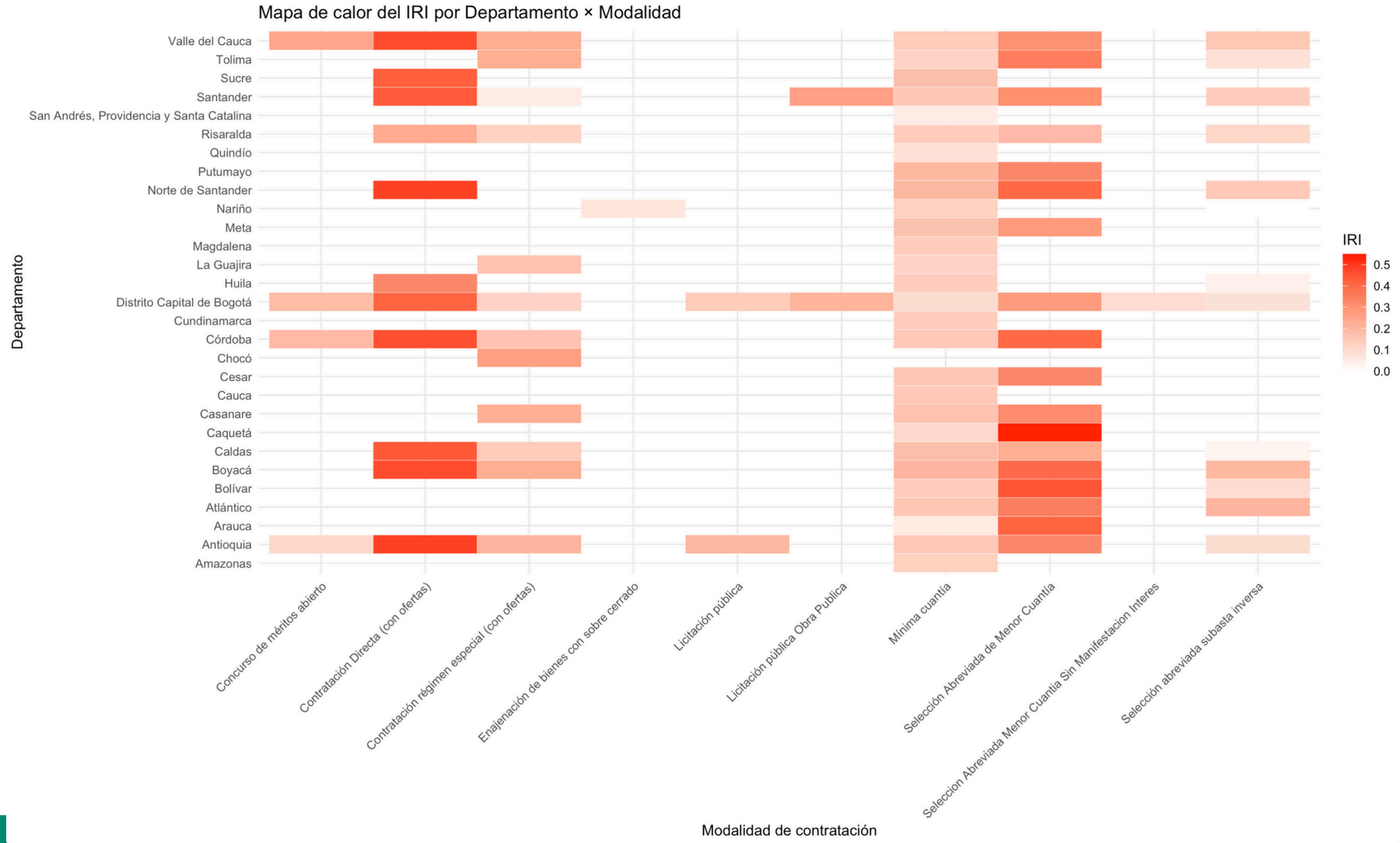


Fuente: SECOP 2020–2024 | Estimación propia

Graficos y tablas de resultados



Graficos y tablas de resultados

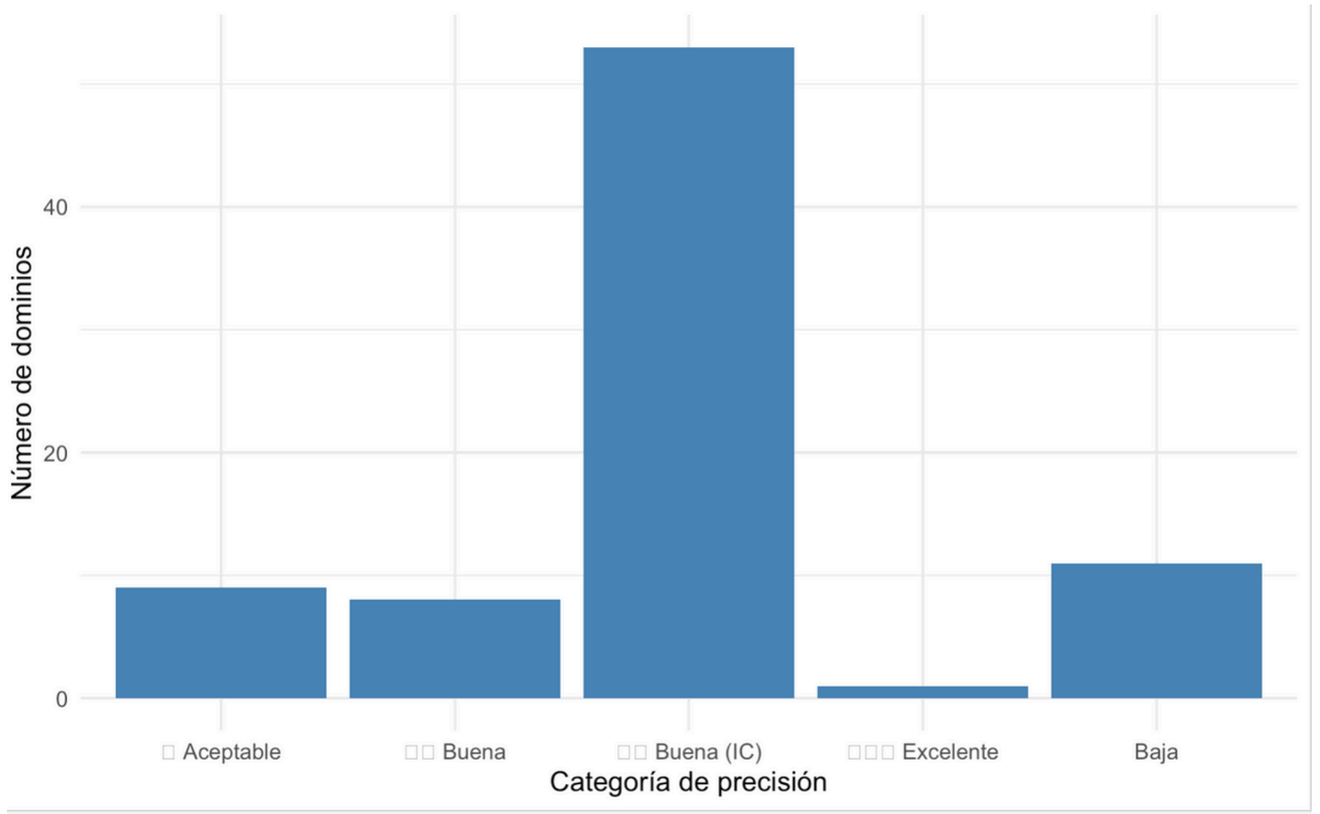


Graficos y tablas de resultados

Distribución de tipos de precisión por dominio

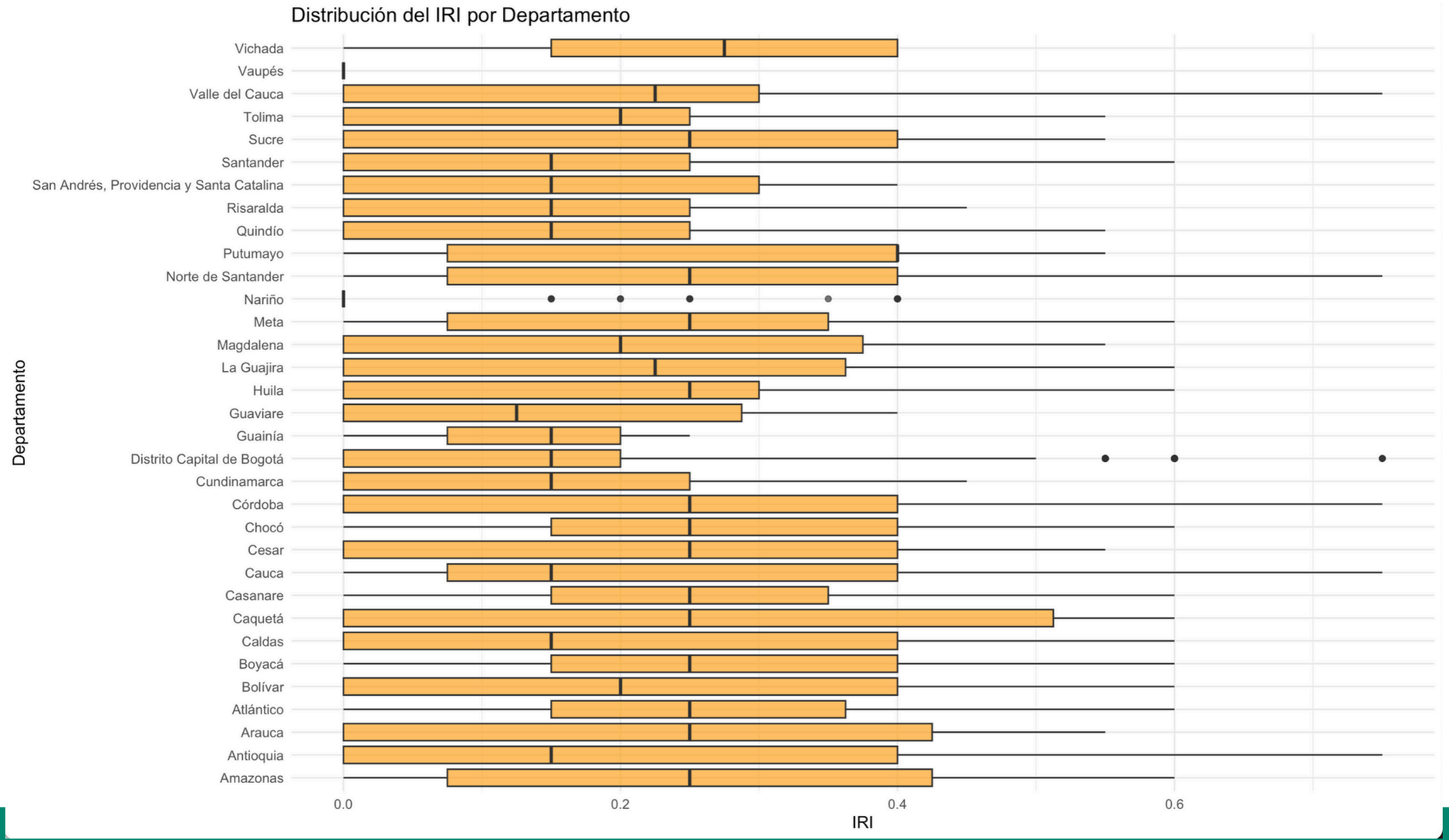
Var1	Freq
★ Aceptable	9
★★ Buena	8
★★ Buena (IC)	53
★★★ Excelente	1
Baja	11

Precisión por dominios



Graficos y tablas de resultados

IRI por Departamento



Graficos y tablas de resultados

IRI por Modalidad