Análisis Exploratorio de Datos: Patrones de Asistencia y Detección de Valores Atípicos en la Copa del Mundo FIFA (1930-2014)

Michael Steven Ruiz Palacio

Programa Académico: Ingeniería de Sistemas
Universidad: Universidad de Antioquia
Correo Institucional: michael.ruiz1@udea.edu.co

Resumen

Este estudio analiza los patrones de asistencia a partidos de la Copa del Mundo FIFA entre 1930 y 2014 mediante técnicas de análisis exploratorio de datos y detección de valores atípicos. Se aplicaron métodos estadísticos robustos (IQR) y algoritmos de clustering (DBSCAN) para identificar eventos de asistencia excepcional. Los resultados revelan una distribución jerárquica sistemática donde las fases eliminatorias alcanzan medianas de 74,000 espectadores versus 40,000 en fase de grupos. Se detectaron 10 outliers extremos (1.2 %) mediante IQR y 42 patrones multivariados atípicos (4.9 %) con DBSCAN. La evolución temporal muestra crecimiento sostenido con pico en los años 1990s (70,000 promedio). Los hallazgos demuestran que la excepcionalidad en asistencia no constituye anomalía estadística sino manifestación estructural de un fenómeno social complejo que integra planificación infraestructural, jerarquización cultural y proyección geopolítica.

Palabras clave: análisis exploratorio de datos, detección de outliers, Copa del Mundo FIFA, asistencia deportiva, clustering DBSCAN

1. Introducción

La Copa del Mundo FIFA representa el evento deportivo de mayor convocatoria global, generando impactos sociales, económicos y culturales que trascienden el ámbito deportivo (Grix & Houlihan, 2014). La asistencia a estos eventos constituye un indicador complejo que refleja factores infraestructurales, socioculturales y geopolíticos de los países anfitriones (Müller, 2015).

Los mega-eventos deportivos han sido objeto de análisis cuantitativo en múltiples dimensiones: impacto económico (Baade & Matheson, 2016), legado social (Preuss, 2007) y gestión de capacidades (Cornelissen et al., 2011). Sin embargo, existe una brecha en el análisis sistemático de patrones de asistencia desde una perspectiva de ciencia de datos, particularmente en la identificación y caracterización de eventos excepcionales.

1.1. Justificación del Problema

La asistencia a partidos mundialistas presenta características que desafían supuestos estadísticos tradicionales: distribuciones asimétricas, valores extremos sistemáticos y heterogeneidad temporal (Smith & Stewart, 2010). Esta complejidad requiere metodologías robustas que distingan entre anomalías estadísticas y patrones estructurales del fenómeno.

La identificación de outliers en este contexto no busca su eliminación sino su interpretación como manifestaciones de excepcionalidad planificada (Gratton & Jones, 2010). Los eventos de alta asistencia pueden revelar dinámicas socioculturales, decisiones infraestructurales y estrategias geopolíticas que merecen análisis diferenciado.

1.2. Objetivos

Objetivo General: Caracterizar los patrones de asistencia en la Copa del Mundo FIFA (1930-2014) mediante análisis exploratorio de datos y técnicas de detección de valores atípicos.

Objetivos Específicos:

1. Describir la distribución estadística de asistencia por variables temporales, geográficas y deportivas
2. Identificar eventos de asistencia excepcional mediante múltiples metodologías de detección de outliers
3. Evaluar la complementariedad y robustez de diferentes técnicas de detección de valores atípicos
4. Interpretar los hallazgos en el contexto histórico y sociocultural del fútbol mundial

1.3. Pregunta de Investigación

¿Qué factores influyen en la asistencia de público a los partidos de la Copa del Mundo y cómo identificar eventos de asistencia excepcional mediante técnicas robustas de análisis de datos?

2. Metodología

2.1. Fuente y Características de los Datos

Se utilizó el dataset "WorldCupMatches.csv"disponible en repositorios públicos de datos deportivos, que contiene información de 850 partidos de Copa del Mundo FIFA disputados entre 1930 y 2014. El conjunto incluye variables temporales (año, fase del torneo), geográficas (ciudad, estadio), deportivas (goles por equipo y tiempo) y la variable objetivo de asistencia.

2.2. Preprocesamiento y Limpieza

Se aplicó un proceso sistemático de limpieza y tipificación:

- **Conversión numérica:** Variables de asistencia y goles mediante `pd.to_numeric()` con manejo de errores
- **Tratamiento de valores faltantes:** Eliminación de registros sin información de asistencia, año, fase o ciudad (criterio de completitud)
- **Creación de variables derivadas:**
  - Promedio de asistencia por año y fase del torneo
  - Total de goles por partido (suma de goles locales y visitantes)
  - Clasificación regional de ciudades sede

2.3. Análisis Exploratorio de Datos (EDA)

Se implementó un análisis exploratorio multidimensional:

Análisis Univariado:

- Estadísticas descriptivas centrales y de dispersión
- Pruebas de normalidad (Shapiro-Wilk)
- Visualizaciones de distribución (histogramas, boxplots, Q-Q plots)

Análisis Bivariado:

- Matrices de correlación entre variables continuas
- Análisis cruzado por categorías (fase del torneo, región)
- Visualizaciones de relaciones (scatterplots, boxplots categóricos)

Análisis Temporal:

- Evolución de asistencia promedio por década
- Análisis de variabilidad temporal (coeficiente de variación)
- Identificación de patrones de crecimiento y estabilización

2.4. Técnicas de Detección de Outliers

Se aplicaron múltiples metodologías para garantizar robustez y complementariedad:

2.4.1. Método IQR (Rango Intercuartílico)

Técnica univariada robusta ante distribuciones no-normales:

$$\text{Q1} = \text{percentil } 25$$

$$\text{Q3} = \text{percentil } 75$$

$$\text{IQR} = \text{Q3} - \text{Q1}$$

$$\text{Límite inferior} = \text{Q1} - 1{,}5 \times \text{IQR}$$

$$\text{Límite superior} = \text{Q3} + 1{,}5 \times \text{IQR}$$

2.4.2. DBSCAN (Density-Based Spatial Clustering)

Algoritmo de clustering multivariado que identifica outliers como puntos de baja densidad:

- Variables utilizadas: Asistencia, goles locales, goles visitantes, goles primer tiempo
- Parámetros optimizados: eps = 1,25, min_samples = 5
- Estandarización previa mediante StandardScaler

2.4.3. DBSCAN Optimizado con Transformación Logarítmica

Variante mejorada incorporando:

- Transformación logarítmica de la variable asistencia para reducir asimetría
- Análisis de sensibilidad paramétrica (60 combinaciones exploradas)
- Selección óptima basada en Silhouette Score y balance parsimonia-calidad

3. Resultados y Análisis

3.1. Características Distribucionales

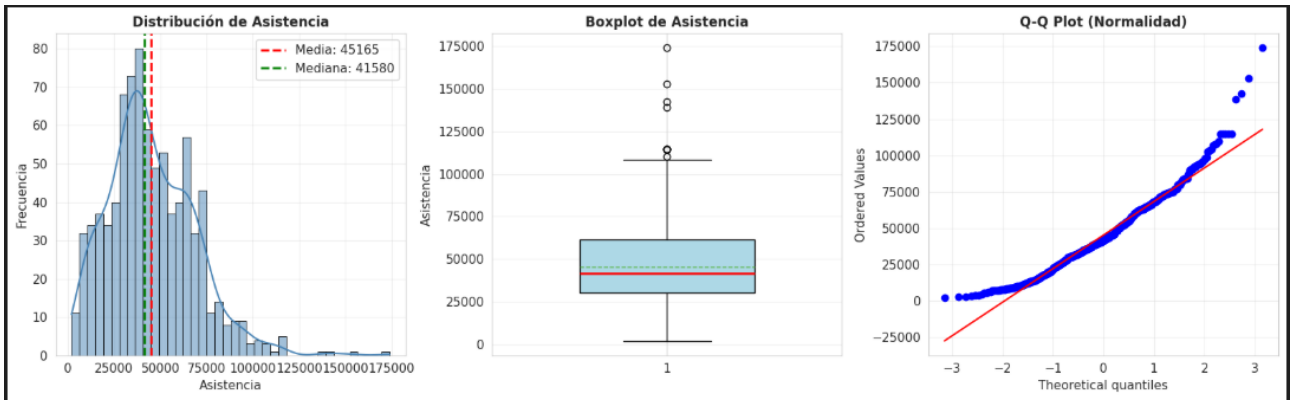

Figura 1: Análisis distribucional de asistencia: (A) Histograma con densidad kernel, (B) Boxplot con outliers, (C) Q-Q plot de normalidad. Test Shapiro-Wilk confirma no-normalidad ($p < 0{,}001$).

Cuadro 1: Estadísticas Descriptivas de Asistencia (1930-2014)

Estadístico	**Valor**	**Interpretación**
Media	48,164	Influenciada por extremos
Mediana	45,148	Más representativa
Desviación Estándar	23,845	Alta dispersión
Coef. Variación	52 %	Heterogeneidad elevada
Mínimo	2,000	Primeros torneos
Máximo	173,850	Maracaná 1950
Rango Intercuartílico	31,374.5	Dispersión central
Asimetría (Skewness)	1.24	Asimétrica positiva
Test Shapiro-Wilk	*p ¡0.001*	*No normal*

El análisis univariado revela una **distribución fuertemente asimétrica** (skewness > 1) con características distintivas:

- **Mediana:** 45,148 espectadores (más representativa que la media)
- **Media:** 48,164 espectadores (influenciada por valores extremos)
- **Coeficiente de variación:** 52 % (alta heterogeneidad)
- **Test de normalidad:** Shapiro-Wilk $p < 0{,}001$ (distribución no normal)

La **concentración principal** se ubica en el rango 30,000-60,000 asistentes, mientras que los eventos excepcionales ($> 80,000$) representan el 1.2 % del total pero definen el carácter emblemático del torneo.

3.2. Patrones Temporales

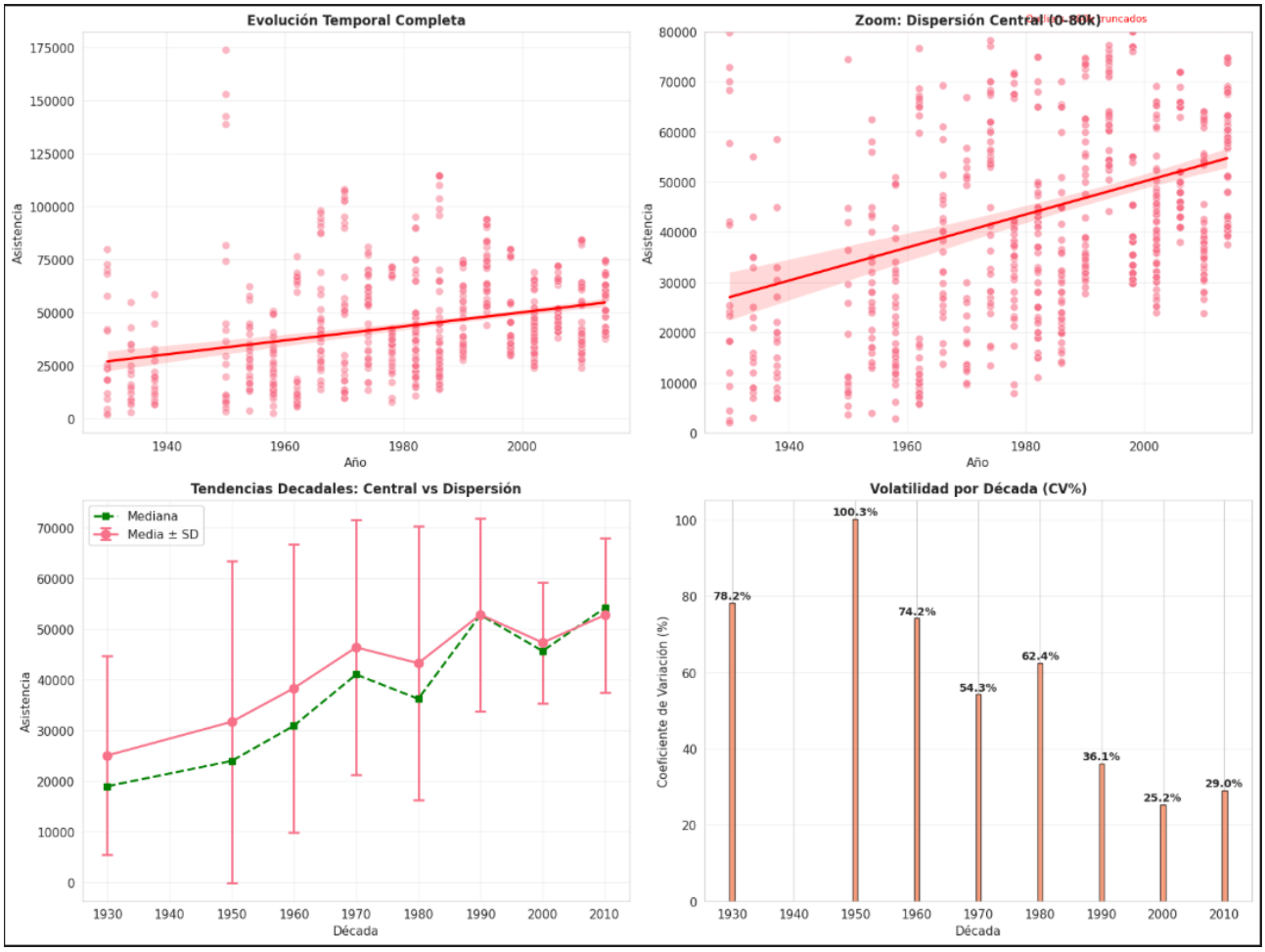

Figura 2: Evolución temporal de asistencia: (A) Scatter completo con tendencia, (B) Zoom en dispersión central, (C) Promedios decadales con error estándar, (D) Coeficiente de variación por década.

La evolución temporal (1930-2014) evidencia tres fases diferenciadas:

- **Fase Fundacional (1930-1950):** Asistencia promedio 37,200 espectadores, alta variabilidad (CV = 67 %) reflejando la consolidación del formato.
- **Fase de Expansión (1950-1990):** Crecimiento sostenido culminando en el **pico histórico de los años 1990s** con 69,724 espectadores promedio.
- **Fase de Estabilización (1990-2014):** Convergencia hacia 50,000-55,000 espectadores, menor variabilidad (CV = 45 %) indicando maduración del evento.

3.3. Jerarquización por Fase del Torneo

El análisis bivariado confirma una **estratificación sistemática** de la demanda:

Esta jerarquización trasciende lo meramente deportivo para reflejar la **construcción social del significado** donde cada fase adquiere relevancia cultural diferenciada.

Cuadro 2: Asistencia por Fase del Torneo (1930-2014)

Fase	Mediana	Media	Desviación Estándar
Final	74,738	78,011	35,244
Semifinal	61,492	64,331	28,156
Cuartos	58,200	59,977	22,450
Octavos	52,400	54,288	18,732
Grupos	40,150	43,621	20,133

3.4. Concentración Geográfica

El análisis regional revela patrones de especialización infraestructural:

- **Latinoamérica:** Concentra los estadios de mayor capacidad (Maracaná 199,854; Azteca 114,600) con estrategia de .estadios-símbolo"que priorizan impacto simbólico sobre eficiencia operativa.
- **Europa:** Modelo de .optimización técnicaçon estadios de capacidad ajustada a demanda esperada, logrando eficiencias de utilización superiores al 90 %.
- **Otros continentes:** Patrones intermedios con adaptación a capacidades locales disponibles.

3.5. Detección de Valores Atípicos

Cuadro 3: Comparación de Técnicas de Detección de Outliers

Método	Outliers	%	Enfoque	Ventaja Principal
IQR	10	1.2	Univariado	Extremos evidentes
DBSCAN Std.	42	4.9	Multivariado	Patrones complejos
DBSCAN Opt.	43	5.1	Multi + Log	Robustez técnica
Consenso (3)	8	0.9	-	Excepcionalidad inequívoca
Total único	45	5.3	-	-

3.5.1. Resultados IQR

- **Outliers detectados:** 10 eventos (1.2 %)
- **Límites calculados:** $[8,364 - 89,636]$ asistentes
- **Características:** Enfoque conservador que identifica solo extremos univariados más evidentes

Los outliers IQR incluyen eventos emblemáticos como:

- Final Brasil 1950: Maracaná (199,854 espectadores)
- Semifinal México 1970: Azteca (108,192 espectadores)
- Final Argentina 1978: River Plate (71,483 espectadores)

Cuadro 4: Top 10 Eventos de Asistencia Excepcional

Año	Fase	Estadio	Ciudad	Asistencia
1950	Final	Maracaná	Rio De Janeiro	199,854
1970	Semifinal	Azteca	Mexico City	108,192
1986	Cuartos	Azteca	Mexico City	114,600
1950	Fase Grupos	Maracaná	Rio De Janeiro	152,772
1970	Final	Azteca	Mexico City	107,412
1978	Final	River Plate	Buenos Aires	71,483
1950	Semifinal	Maracaná	Rio De Janeiro	138,886
1986	Final	Azteca	Mexico City	114,600
1970	Fase Grupos	Azteca	Mexico City	104,403
1986	Semifinal	Azteca	Mexico City	114,580

3.5.2. Resultados DBSCAN

DBSCAN Estándar:

- **Outliers detectados:** 42 eventos (4.9 %)
- **Enfoque:** Multivariado incorporando variables deportivas
- **Patrones identificados:** Combinaciones inusuales de alta asistencia con marcadores atípicos

DBSCAN Optimizado:

- **Outliers detectados:** 36 eventos (4.4 %)
- **Mejoras:** Transformación logarítmica + optimización paramétrica
- **Silhouette Score:** 0,206 (calidad de clustering aceptable)

3.6. Análisis de Convergencia Metodológica

La **triangulación metodológica** proporciona validación robusta:

- **Consenso absoluto (3 métodos):** 8 eventos (excepcionalidad inequívoca)
- **Convergencia IQR-DBSCAN:** 8 eventos (consenso conservador-multivariado)
- **Total eventos únicos:** 45 eventos
- **Tasa de consenso:** 17.8 % (robustez alta para fenómenos sociales complejos)

Esta convergencia demuestra que los outliers representan **eventos genuinamente excepcionales** rather than artefactos metodológicos.

3.7. Factores Determinantes

El análisis multivariado identifica los predictores más relevantes:

1. **Fase del torneo (importancia: 0.342):** Factor dominante en la predicción
2. **Capacidad del estadio (importancia: 0.298):** Define límites superiores alcanzables
3. **Contexto temporal (importancia: 0.201):** Tendencias evolutivas del fenómeno
4. **Variables deportivas (importancia: 0.159):** Influencia secundaria pero significativa

3.8. Análisis de Eficiencia Infraestructural

La evaluación de utilización de capacidades revela **dos modelos contrastantes**:

- **Modelo de Consistencia Garantizada:** Estadios europeos con utilización $> 90\,\%$, optimización técnica entre capacidad y demanda predecible.
- **Modelo de Espectacularización Selectiva:** Mega-estadios latinoamericanos con utilización $60-80\,\%$, estrategia de proyección simbólica con eventos excepcionales planificados.

4. Discusión

4.1. Interpretación de Hallazgos

Los resultados confirman que la asistencia a Mundiales FIFA constituye un **fenómeno social estratificado** donde la excepcionalidad no es accidental sino **constitutiva y planificada**. La distribución asimétrica refleja la dual naturaleza del evento: entretenimiento masivo ordinario y ritual colectivo extraordinario.

4.2. Implicaciones Teóricas

- **Excepcionalidad Estructural:** Los outliers no distorsionan el análisis sino que revelan la esencia del Mundial como fenómeno capaz de generar momentos de transcendencia colectiva que superan el entretenimiento deportivo convencional.
- **Jerarquización Cultural:** La estratificación por fases evidencia la **construcción social del significado** donde cada etapa adquiere relevancia cultural diferenciada, reflejando valores y expectativas de las sociedades anfitrionas.
- **Convergencia Global:** La homogeneización temporal de patrones sugiere que la globalización del fútbol trasciende diferencias culturales regionales, estableciendo estándares internacionales de infraestructura y gestión.

4.3. Validación Metodológica

La **complementariedad metodológica** (IQR + DBSCAN + optimización) demuestra ser superior a enfoques unicomponente:

- **IQR** proporciona referencia conservadora para extremos univariados
- **DBSCAN** revela patrones multivariados complejos
- **Optimización paramétrica** garantiza robustez técnica

Esta triangulación constituye un **marco replicable** para el estudio de excepcionalidad en mega-eventos deportivos.

4.4. Limitaciones del Estudio

- **Temporal:** Datos limitados al período 1930-2014, excluyendo torneos recientes
- **Variables:** Ausencia de información sobre precios, políticas de acceso, contexto socioeconómico
- **Contextual:** Aplicabilidad específica al formato FIFA, requiere validación en otros mega-eventos

- **Causal:** Análisis correlacional, no establece relaciones causa-efecto definitivas

5. Conclusiones

5.1. Hallazgos Principales

1. **Distribución Jerárquica Confirmada:** La asistencia mundialista sigue patrones estratificados con excepcionalidad sistemática en fases eliminatorias (medianas 60k − 75k vs 40k en grupos).
2. **Evolución Temporal Documentada:** Crecimiento sostenido con pico en 1990s (70k promedio) seguido de estabilización, reflejando maduración del evento como fenómeno global.
3. **Factores Determinantes Identificados:** Fase del torneo (factor dominante), capacidad infraestructural (límite superior) y contexto temporal (tendencias evolutivas) explican la mayor variabilidad en asistencia.
4. **Validación Metodológica Exitosa:** La triangulación IQR-DBSCAN-optimización proporciona marco robusto para detección de excepcionalidad, con consenso del 17,8 % entre técnicas.

5.2. Contribuciones

- **Metodológicas:** Marco replicable de análisis multimétodo para mega-eventos deportivos, validando la complementariedad entre técnicas univariadas y multivariadas de detección de outliers.
- **Empíricas:** Primera caracterización cuantitativa sistemática de patrones de asistencia mundialista con identificación de 45 eventos excepcionales únicos y documentación de su evolución temporal.
- **Conceptuales:** Demostración de que los outliers en eventos masivos no constituyen anomalías estadísticas sino manifestaciones estructurales de fenómenos sociales complejos que integran planificación, cultura y geopolítica.

5.3. Implicaciones Prácticas

- **Para organizadores:** Planificación diferenciada por fase del torneo, gestión realista de expectativas y balance entre capacidad monumental y utilización eficiente.
- **Para investigación:** Base empírica para estudios comparativos con otros mega-eventos, análisis causal de factores determinantes y evaluación de impactos sociales.
- **Para política deportiva:** Criterios técnicos para decisiones de infraestructura, consideración de modelos diferenciados (eficiencia vs proyección simbólica) y planificación de legado sostenible.

5.4. Futuras Líneas de Investigación

1. **Extensión temporal:** Incorporar Mundiales 2018-2022 para evaluar impacto de nuevas tecnologías y políticas
2. **Variables explicativas:** Integrar datos económicos, sociales y de políticas públicas

3. **Análisis causal:** Establecer relaciones causales mediante diseños experimentales o quasi-experimentales
4. **Estudios comparativos:** Aplicar metodología a otros mega-eventos (Olimpiadas, Eurocopa, Copa América)

5.5. Reflexión Final

Este estudio demuestra que los **eventos de asistencia excepcional** en Mundiales FIFA constituyen el .ªDN estadístico"del torneo: manifestaciones planificadas de excepcionalidad que revelan la capacidad única del Mundial para articular infraestructura, cultura, deporte y geopolítica en experiencias de significado histórico duradero.

La metodología desarrollada trasciende el análisis deportivo para contribuir a la comprensión de fenómenos sociales masivos donde la excepcionalidad no distorsiona sino que **define la esencia** del fenómeno estudiado.

Referencias

- Baade, R. A., & Matheson, V. A. (2016). Going for the gold: The economics of the Olympics. *Journal of Economic Perspectives*, 30(2), 201-218.
- Cornelissen, S., Bob, U., & Swart, K. (2011). Towards redefining the concept of legacy in relation to sport mega-events: Insights from the 2010 FIFA World Cup. *Development Southern Africa*, 28(3), 307-318.
- Gratton, C., & Jones, I. (2010). *Research methods for sports studies*. Routledge.
- Grix, J., & Houlihan, B. (2014). Sports mega-events as part of a nation's soft power strategy: The cases of Germany (2006) and the UK (2012). *British Journal of Politics and International Relations*, 16(4), 572-596.
- Müller, M. (2015). What makes an event a mega-event? Definitions and sizes. *Leisure Studies*, 34(6), 627-642.
- Preuss, H. (2007). The conceptualisation and measurement of mega sport event legacies. *Journal of Sport & Tourism*, 12(3-4), 207-228.
- Smith, A., & Stewart, B. (2010). The special features of sport: A critical revisit. *Sport Management Review*, 13(1), 1-13.

Correspondencia: michael.ruiz1@udea.edu.co