

Resumen Esquemático de Teoría para Visualización de Datos

I. Introducción a la Teoría de la Visualización de Datos

El presente informe ofrece un resumen esquemático y conciso de los conceptos teóricos fundamentales en Visualización de Datos. Su propósito es facilitar la revisión eficiente de definiciones clave, metodologías y mejores prácticas, abordando específicamente el tipo de preguntas teóricas que se presentan en los exámenes académicos. El alcance de este compendio abarca el procesamiento de datos, la reducción de dimensionalidad, la visualización de múltiples variables, la teoría del color, la precisión visual, la interactividad, la animación y la experiencia de usuario (UX).

II. Procesamiento de Datos y Reducción de Dimensionalidad

El procesamiento de datos constituye un paso crítico para una visualización efectiva, implicando a menudo transformaciones y manipulación de datos. Un aspecto central de este proceso es la reducción de dimensionalidad, que aborda el desafío de visualizar conjuntos de datos con un elevado número de atributos.

Reducción de Dimensionalidad (DR): Definición y Objetivos

La Reducción de Dimensionalidad es el proceso de transformar datos de un espacio de alta dimensionalidad a uno de baja dimensionalidad, conservando la mayor parte de la información significativa del conjunto de datos original.¹ Su objetivo principal es mantener la estructura relevante del conjunto de datos utilizando un menor número

de atributos.¹ Por ejemplo, esto podría implicar la reducción de cientos de columnas a una representación bidimensional o tridimensional.¹

La DR persigue dos objetivos fundamentales: la exploración de datos y el aprendizaje automático.¹ En el ámbito de la visualización, gestiona la complejidad al permitir que las muestras se visualicen en menos dimensiones (por ejemplo, 2D o 3D), lo que facilita la obtención de conocimientos, la detección de clústeres y la identificación de valores atípicos.¹ La DR se puede lograr mediante la selección de características (reduciendo el espacio de características al eliminar o seleccionar las necesarias) o mediante métodos basados en componentes/factores y en proyección.¹

Técnicas Clave de Reducción de Dimensionalidad

Análisis de Componentes Principales (PCA):

El PCA introduce un nuevo y más pequeño conjunto de variables, denominadas componentes principales (PCs), que son combinaciones lineales de las variables originales. Estas variables originales suelen estandarizarse a media cero y varianza unitaria.¹ Es un algoritmo no supervisado cuyo objetivo es encontrar ejes de componentes ortogonales que maximicen la varianza en un conjunto de datos.¹ El PCA es adecuado cuando se necesita reducir el número de variables, pero no se pueden identificar variables específicas para su eliminación. También asegura que las variables sean independientes entre sí, aunque esto puede hacer que las variables independientes sean menos interpretables.¹

Los conceptos fundamentales del PCA incluyen:

- **Varianza:** Representa la variación de los valores dentro de una única variable, indicando cuán dispersos están los valores entre sí.¹
- **Covarianza:** Mide cómo dos variables diferentes varían conjuntamente. A diferencia de la varianza, indica la dirección de su relación lineal.¹
- **Matriz de Covarianza:** Una matriz simétrica que representa los valores de covarianza para cada par de variables en datos multivariados. Los elementos de la diagonal muestran la varianza de cada variable.¹
- **Valores y Vectores Propios:** Para la matriz de covarianza, los valores propios representan la magnitud de la dispersión en la dirección de las componentes principales, mientras que los vectores propios indican la dirección de estas componentes.¹ La primera PC (PC1) representa la dirección de mayor varianza, y las PCs subsiguientes capturan la mayor varianza restante ortogonal a las anteriores.¹

El PCA es altamente interpretable, ya que cada PC está bien definida y es ortogonal. Permite cuantificar la varianza explicada por cada PC para seleccionar un número apropiado de dimensiones.¹ Sin embargo, el PCA es muy sensible a los valores atípicos y funciona mejor solo con datos continuos.¹

Análisis Discriminante Lineal (LDA):

El LDA es un algoritmo supervisado que busca la mejor manera de separar (o discriminar) las muestras en el conjunto de datos de entrenamiento según su valor de clase.¹ Su objetivo es encontrar un nuevo espacio de características que maximice la separabilidad entre clases.¹ Aunque tanto el PCA como el LDA son técnicas de transformación lineal para la reducción de dimensionalidad, sus objetivos difieren: el PCA se centra en maximizar la varianza en los datos (no supervisado), mientras que el LDA busca optimizar la separabilidad de clases (supervisado).¹ El LDA no funciona bien si el diseño del conjunto de datos está desequilibrado (número desigual de objetos en las clases) o si la distribución de los datos no es significativamente gaussiana. También puede ser sensible al sobreajuste y no es aplicable para problemas no lineales.¹

T-Distributed Stochastic Neighbor Embedding (t-SNE):

El t-SNE es una técnica de reducción de dimensionalidad no lineal utilizada principalmente para visualizar conjuntos de datos de alta dimensionalidad. Asigna a cada punto de datos una ubicación en un mapa 2D o 3D, preservando la estructura local de los datos.¹ A diferencia del PCA, el t-SNE se enfoca en la deconvolución de las relaciones entre vecinos en datos de alta dimensionalidad, lo que lo hace efectivo para identificar clústeres locales. El PCA, al ser lineal, puede distorsionar las distancias en conjuntos de datos altamente no lineales.¹ Las limitaciones del t-SNE incluyen su alto costo computacional, su naturaleza probabilística (los resultados pueden variar entre ejecuciones) y el hecho de que no preserva la estructura global entre clústeres (las distancias entre clústeres carecen de significado).¹

Correlogramas: Propósito e Interpretación

Los correlogramas visualizan la estructura de correlación de las variables de entrada.¹ La intensidad del color y la forma de la elipse están directamente relacionadas con el coeficiente de correlación (R).¹ Las correlaciones positivas ($R > 0$) suelen mostrarse en azul, mientras que las negativas ($R < 0$) se representan en rojo. Una mayor intensidad de color y una elipse más estrecha indican una correlación más fuerte.¹ Los correlogramas son particularmente útiles cuando las matrices de diagramas de dispersión (SPLOM) resultan inmanejables debido a un gran número de variables cuantitativas (>3 o 4), ya que cuantifican y visualizan la asociación entre pares de variables de manera más efectiva.¹

Tabla 2: Conceptos Clave en Reducción de Dimensionalidad

Técnica	Definición	Propósito Principal	Linealidad	Supervisa do/No Supervisa do	Fortalezas	Debilidade s
PCA	Transform a datos a un nuevo conjunto de variables ortogonales (PCs)	Maximizar la varianza y reducir dimensiones	Lineal	No Supervisa do	Muy interpretable, PCs ortogonales, cuantifica varianza explicada	Afectado por valores atípicos, mejor con datos continuos
LDA	Proyecta datos para maximizar la separación entre clases	Optimizar la separabilidad de clases	Lineal	Supervisa do	Efectivo para clasificación, maximiza la discriminación	No funciona bien con datos desequilibrados o no gaussianos, sensible al sobreajuste
t-SNE	Reduce dimensiones preservando la estructura local de los datos	Visualizar conjuntos de datos de alta dimensionalidad, identificar clústeres locales	No Lineal	No Supervisa do	Revela estructuras de clústeres complejas en datos no lineales	Costoso computacionalmente, probabilístico, no preserva la estructura global entre clústeres

La efectividad de las técnicas de reducción de dimensionalidad no depende únicamente del algoritmo en sí, sino que está fundamentalmente influenciada por la calidad y las características de los datos de entrada. Por ejemplo, el PCA se ve "altamente afectado por los valores atípicos" y funciona mejor con "datos continuos".¹

De manera similar, el LDA es sensible a distribuciones "no gaussianas" y diseños "desequilibrados".¹ Esto pone de manifiesto que un preprocesamiento de datos robusto, que incluya la detección y el manejo de valores atípicos, la normalización de datos y la gestión de valores faltantes, no es simplemente un paso preliminar, sino una parte integral del proceso de reducción de dimensionalidad. Una calidad de datos deficiente en esta etapa puede conducir a proyecciones engañosas, lo que a su vez hace que los conocimientos derivados de la visualización sean poco fiables. Esta interdependencia subraya la importancia de la fiabilidad y la estabilidad de los datos en todo el ciclo de vida de la visualización.

La elección de una técnica de reducción de dimensionalidad implica una compensación entre la interpretabilidad y la capacidad de revelar patrones complejos. El PCA se describe como "muy interpretable" debido a la ortogonalidad de sus componentes y la cuantificación de la varianza explicada.¹ Por el contrario, el t-SNE es "computacionalmente costoso", "probabilístico" y "no preserva la estructura global entre clústeres".¹ Esta distinción implica que la selección de la técnica no se limita a la mera reducción de dimensiones, sino que busca un equilibrio entre la necesidad de una interpretabilidad clara, por ejemplo, para explicar la importancia de las características en un modelo, y la capacidad de descubrir patrones no lineales complejos, como en el análisis exploratorio de datos para encontrar clústeres ocultos. Para un informe técnico, la claridad del PCA podría ser preferible, mientras que para la obtención de conocimientos exploratorios, el t-SNE podría ser más potente a pesar de sus complejidades inherentes.

III. Visualización de Múltiples Variables y Dimensiones

La visualización de conjuntos de datos con numerosas variables y dimensiones representa un desafío significativo, que requiere técnicas especializadas más allá de los simples diagramas de dispersión o gráficos de barras.

Enfoques Generales para Múltiples Variables y Dimensiones

Los conjuntos de datos grandes a menudo contienen más información de la que se

puede mostrar en un solo gráfico.¹ Para conjuntos de datos complejos, las figuras de paneles múltiples pueden ser útiles, consistiendo en varios paneles que muestran subconjuntos de datos.¹

Las **Múltiples Pequeñas (Small Multiples)**, también conocidas como "gráficos de enrejado" o "facetar", son gráficos dispuestos en una cuadrícula regular. Cada panel muestra un subconjunto diferente de los datos, pero todos los paneles utilizan el mismo tipo de visualización.¹ La idea central es dividir los datos por una o más dimensiones, visualizar cada segmento por separado y organizarlos en una cuadrícula. Es fundamental evitar diferentes escalas en los ejes entre paneles y organizarlos en un orden lógico y significativo.¹

Las **Figuras Compuestas** combinan varios paneles independientes en una única figura para transmitir un punto general. A diferencia de las múltiples pequeñas, estos paneles pueden organizarse de forma arbitraria y mostrar visualizaciones o conjuntos de datos completamente diferentes. Es crucial mantener la coherencia en colores, símbolos y fuentes en todos los paneles, y asegurar que los ejes estén alineados para evitar interpretaciones erróneas.¹

Visualización de Múltiples Distribuciones a la Vez

Este enfoque implica mostrar la distribución de una "variable de respuesta" a través de diferentes "variables de agrupación".¹ Ejemplos de gráficos adecuados para este propósito incluyen diagramas de caja (Boxplots), diagramas de violín (Violin Plots), gráficos de franjas (Strip Charts), histogramas apilados (Stacked Histograms), densidades superpuestas (Overlapping Densities), gráficos de cresta (Ridgeline Plots) y gráficos de Sina (Sina Plots).¹

Visualización de Múltiples Proporciones a la Vez

Este apartado aborda escenarios en los que las proporciones se especifican según múltiples variables de agrupación o cambian a lo largo de una variable continua.¹ Mientras que los gráficos circulares múltiples (Multiple Pie Charts) son ineficientes en el uso del espacio, los gráficos de barras agrupadas (Grouped Bars) funcionan bien

para un número moderado de condiciones, y los gráficos de barras apiladas (Stacked Bars) son adecuados para un gran número.¹ Las densidades apiladas (Stacked Densities) son apropiadas cuando las proporciones cambian a lo largo de una variable continua.¹

Treemaps:

Un Treemap es un gráfico rectangular dividido en mosaicos, donde cada mosaico representa una única observación. Muestra datos jerárquicos utilizando rectángulos anidados.¹ El área relativa de cada mosaico expresa una variable continua.¹ Los Treemaps funcionan bien incluso si las subdivisiones de un grupo son completamente distintas de las subdivisiones de otro.¹ Su construcción implica anidar rectángulos de forma recursiva y requiere una variable numérica continua para el área del mosaico, una variable para el color de relleno (grupo padre) y una para la etiqueta del mosaico.¹

Gráficos de Mosaico (Mosaic Plots):

Los gráficos de mosaico son similares a los gráficos de barras apiladas, pero tanto la altura como la anchura de las áreas sombreadas individuales varían. Estas áreas son proporcionales al número de casos que representan cada posible combinación de dos variables categóricas.¹ Un gráfico de mosaico asume que cada nivel de una variable de agrupación puede combinarse con cada nivel de otra variable de agrupación.¹ A diferencia de los gráficos de barras apiladas, donde normalmente solo varía la altura, los gráficos de mosaico varían tanto la altura como la anchura.¹

Conjuntos Paralelos (Parallel Sets):

Los Conjuntos Paralelos son una técnica que muestra cómo el conjunto de datos total se desglosa por cada variable categórica individual, y luego dibuja bandas sombreadas para mostrar cómo se relacionan los subgrupos entre sí.¹ Esta técnica funciona mejor que los gráficos de mosaico o los treemaps cuando hay más de dos variables de agrupación.¹ Los Conjuntos Paralelos son superiores para visualizar proporciones descritas por más de dos variables categóricas, mientras que los treemaps y los gráficos de mosaico suelen limitarse a una o dos variables de agrupación.¹

Visualización de Múltiples Relaciones (Correlaciones) a la Vez

Matriz de Diagramas de Dispersión (SPLOM):

Una SPLOM utiliza múltiples diagramas de dispersión para determinar la correlación (si la hay) entre una serie de variables. Estos diagramas de dispersión se organizan en una matriz para facilitar la visualización de posibles correlaciones.¹ Puede identificar relaciones por pares, su naturaleza, valores atípicos y agrupaciones.¹ Sin embargo, las SPLOM se vuelven rápidamente inmanejables cuando hay más de 3 o 4 variables cuantitativas.¹ En tales casos, cuantificar las asociaciones con coeficientes de correlación y visualizar estas cantidades (por ejemplo, con correlogramas) es más efectivo.¹

Gráficos de Burbujas (Bubble Charts):

Los gráficos de burbujas representan datos utilizando círculos u otras formas donde el tamaño y el color del punto codifican variables adicionales.¹ Son ideales para graficar la relación entre tres atributos cuantitativos (dos para la posición, uno para el tamaño) y pueden escalar bien a cientos de observaciones.¹ Son adecuados cuando una tercera variable cuantitativa necesita ser representada visualmente, como la población en un gráfico de PIB frente a esperanza de vida.¹

Tabla 1: Comparación de Técnicas de Visualización Multivariable

Técnica	Definición	Uso Principal	Variables de Agrupación	Ventajas	Desventajas
Treemap	Rectángulos anidados donde el área representa un valor	Datos jerárquicos, proporcionen en una estructura de árbol	1 (para área), 1-2 (para anidamiento/ color)	Eficaz para mostrar jerarquías y proporcionen s relativas, buen uso del espacio	Puede ser difícil comparar tamaños de rectángulos pequeños, la jerarquía profunda puede ser compleja
Gráfico de Mosaico	Rectángulos donde ancho y alto varían según proporcionen s	Proporcionen s entre dos variables categóricas	2	Muestra la relación entre dos variables categóricas, fácil de ver combinacionen s	Asume que todos los niveles de una variable pueden combinarse con todos los de otra, puede ser complejo con muchas categorías
Conjuntos Paralelos	Bandas sombreadas que conectan categorías a través de ejes paralelos	Proporcionen s entre más de dos variables categóricas	>2	Muestra relaciones complejas entre múltiples variables categóricas, permite seguir flujos	Puede generar "spaghetti plot" con muchos datos, difícil de ver patrones claros sin

					interacción
--	--	--	--	--	-------------

El desafío fundamental de la "maldición de la dimensionalidad" en la visualización se manifiesta claramente en la dificultad de las SPLOM para manejar más de 3 o 4 variables ¹, lo que impulsa la necesidad de técnicas de reducción de dimensionalidad o gráficos multivariantes especializados como los Treemaps y los Conjuntos Paralelos. Esto subraya que la elección de la técnica de visualización no es arbitraria, sino una consecuencia directa de la dimensionalidad de los datos. Para datos de alta dimensionalidad, las técnicas que reducen las dimensiones o agregan/anidan información no son solo opciones, sino necesidades para mantener la interpretabilidad y evitar representaciones engañosas. Este es un principio central en el diseño de la visualización de datos: adaptar la complejidad a una codificación visual adecuada.

La distinción entre los gráficos de mosaico, que asumen que "cada nivel de una variable de agrupación puede combinarse con cada nivel de otra", y los Treemaps, que funcionan "incluso si las subdivisiones de un grupo son completamente distintas" ¹, así como la preferencia por los Conjuntos Paralelos para "más de dos variables de agrupación" ¹, revela la importancia de la estructura subyacente de los datos. Los Treemaps, por ejemplo, están diseñados explícitamente para "datos jerárquicos".¹ Esto implica que comprender la estructura inherente de las variables categóricas (por ejemplo, si forman una jerarquía estricta, son independientes o tienen relaciones complejas de muchos a muchos) es crucial para seleccionar la visualización más apropiada. Aplicar incorrectamente una técnica, como usar un gráfico de mosaico para categorías no combinables, conduciría a una visualización inexacta o ininteligible. Esto enfatiza la importancia del modelado de datos antes de la visualización.

IV. Teoría del Color en Visualización de Datos

El color es un canal visual potente, pero su uso efectivo y honesto requiere una comprensión profunda de la percepción humana y de las posibles trampas.

Percepción del Color

La percepción humana del color se basa en tres tipos de conos que reaccionan a diferentes espectros de luz, interpretados en términos de luminosidad, tono y saturación. En contraste, los ordenadores procesan bandas de frecuencia estrechas de RGB y calculan la luz de forma lineal.¹ La Teoría del Proceso Oponente explica cómo el cerebro procesa las señales de color mediante sustracción (por ejemplo, canales de diferencia rojo-verde, amarillo-azul) y adición (canal de luminancia). Esto significa que algunos colores, como el amarillo, pueden percibirse con mayor intensidad debido a la sobreexcitación de los receptores.¹ Además, la percepción humana es relativa; el contexto influye en cómo se perciben los colores.¹

Canales de Color: Tono, Saturación, Luminancia (Mapeo a Tipos de Datos)

El color es un canal potente para codificar datos. Propiedades como la luminancia y la saturación sirven como canales de magnitud (para atributos ordenados), y el tono como canal de identidad (para atributos categóricos).¹

- **Tono:** Es el más adecuado para datos categóricos, donde se prefieren grandes diferencias de tono, saturación y/o luminosidad.¹ Los tonos no tienen un orden inherente.¹
- **Saturación:** Es un canal de magnitud, utilizado para atributos ordenados.
- **Luminancia/Claridad/Brillo:** Es un canal de magnitud, crucial para distinguir detalles y crear contraste.¹

Escalas de Color

- **Categóricas/Cualitativas:** Se utilizan para categorías distintas. Deben evitar gradientes continuos que podrían confundirse con valores ordinales. Se recomienda limitar el número de colores a un máximo de 12.¹
- **Secuenciales:** Para datos cuantitativos que progresan de valores bajos a altos. Pueden ser continuas o discretizadas (ordinales).¹ Semánticamente, una gama de azules para la precipitación o verdes para la vegetación son apropiadas.¹
- **Divergentes:** Para datos cuantitativos con un punto medio significativo (por ejemplo, cero). Los colores divergen de un color neutro central (a menudo

blanco). El color neutro debe representar el valor central.¹ Un ejemplo es la temperatura de -20°C a 30°C.¹

- **Esquemas de Color Agrupados:** Para múltiples categorías con subcategorías, se utilizan diferentes tonos para las categorías principales y se varía la saturación/luminosidad para las subcategorías.¹
- **Mapas Coropléticos Bivariados:** Utilizan dos escalas superpuestas en dos direcciones, a menudo resultando en una matriz de 3x3.¹

Problemas con las Escalas Arcoíris

Las escalas arcoíris suelen ser predeterminadas, pero son problemáticas porque no tienen en cuenta la percepción humana del color.¹ Carecen de un orden inherente, dificultan la visualización de detalles y pueden crear límites falsos en los datos.¹ La variación no lineal y la sobreexcitación de los receptores en ciertos puntos rompen la linealidad de la escala, lo que lleva a "bandas perceptuales" y a la pérdida de detalles.¹ Este es un problema teórico recurrente.¹

Escalas de Color Perceptualmente Correctas (ej., Viridis)

Es crucial utilizar escalas perceptualmente correctas que varíen de manera consistente en todo el rango de valores.¹ Viridis es una escala de color perceptualmente correcta recomendada, diseñada para corregir la deficiencia de color y la impresión en blanco y negro. Otras escalas similares incluyen magma, plasma, inferno, cividis, mako, rocket y turbo.¹ Estas se basan en los modelos de color HCL y HSV.¹

Accesibilidad: Diseño para la Deficiencia de Color

Es vital diseñar para la deficiencia de color utilizando simuladores (por ejemplo, Coblis, Color Brewer 2.0, el paquete colorblindness en R) para asegurar que las visualizaciones sean accesibles para individuos con diferentes tipos de daltonismo

(Deuteranopia, Protanopia, Tritanopia).¹

Semántica del Color

Asignar colores basándose en su significado inherente (por ejemplo, paletas cálidas para altas temperaturas, azul-rojo para temperaturas negativas a positivas, verde para la vegetación) mejora la comprensión.¹ Las elecciones de color incorrectas pueden introducir efectos engañosos que no están presentes en los datos.¹

El problema central de las escalas arcoíris radica en la discrepancia entre cómo los ordenadores procesan el color (linealmente en RGB) y cómo los humanos lo perciben (de forma no lineal, con sensibilidades variables a diferentes tonos y sobreexcitación para ciertos colores como el amarillo).¹ Esta "brecha perceptual" conduce a la creación de "límites falsos" y a la "pérdida de detalles".¹ Esto no es meramente un problema estético, sino una cuestión fundamental de honestidad visual e integridad de los datos. Si la propia visualización introduce artefactos u oculta patrones reales debido a sesgos perceptuales, socava directamente el propósito de la visualización de datos, que es transmitir información con precisión. Esto subraya que la elección del color es una decisión científica, no solo artística, especialmente en contextos técnicos.

El uso del color va más allá de la simple estética, convirtiéndose en una herramienta estratégica para guiar la atención y la interpretación. El documento destaca que "un mayor contraste facilita la detección" y que "la variación en el tono y la luminosidad facilita la búsqueda".¹ También se discute el uso del color para un significado "semántico".¹ Esto implica que, al comprender cómo los diferentes canales de color (tono, saturación, luminosidad) se asignan a los tipos de datos y cómo el contraste afecta la percepción, un visualizador puede dirigir intencionadamente la mirada hacia patrones clave, valores atípicos o relaciones, mejorando así la eficiencia de la transferencia de información y apoyando la narrativa general o el objetivo de la visualización.

V. Honestidad y Precisión Visual

Mantener la honestidad y la precisión en la visualización de datos es primordial para evitar interpretaciones erróneas y garantizar la integridad del mensaje transmitido. Esto implica un manejo cuidadoso de los ejes, las agregaciones, los datos faltantes y la selección de métricas.

Manejo de Ejes

Línea Base Cero para Barras: Los gráficos de barras codifican el tamaño/magnitud y, por lo tanto, **deben** tener una línea base cero para representar con precisión la proporción o cantidad real. No hacerlo puede magnificar significativamente pequeñas diferencias y llevar a conclusiones engañosas.¹

Posición para Gráficos de Líneas: Los gráficos de líneas codifican la posición de los puntos y no requieren inherentemente una línea base cero.¹ Sin embargo, la elección del rango del eje aún puede enfatizar la variación o la tendencia.¹

Visualización Completa del Eje: Mostrar el rango completo del eje es crucial para evitar magnificar o minimizar tendencias.¹ Los ejes truncados pueden distorsionar la percepción, haciendo que cambios menores parezcan dramáticos.¹

Etiquetas: Las etiquetas en los ejes son indispensables para la claridad y la comprensión de los datos.¹

Evitar Ejes Duales: Los ejes duales son generalmente problemáticos porque pueden sugerir implícitamente una correlación entre líneas que puede no existir, lo que lleva a conclusiones incorrectas. Aunque son aceptables en casos muy específicos, su potencial de mala interpretación es alto.¹

Agregaciones de Datos: Asegurar la Equivalencia

Al comparar datos agregados, es fundamental asegurar que las agregaciones sean equivalentes (por ejemplo, comparar datos en períodos de tiempo normalizados o per cápita) para evitar conclusiones engañosas.¹

Manejo de Datos Faltantes: Marcado Explícito

Los datos faltantes siempre deben marcarse explícitamente en la visualización. Esto implica mantener el espacio para las marcas faltantes o usar marcadores distintos para indicar su ausencia, en lugar de simplemente omitirlos, lo que puede crear una falsa sensación de continuidad o completitud.¹

Elección de Métricas Relevantes: Tasas vs. Totales

Seleccionar la métrica adecuada es crucial para una representación honesta. Por ejemplo, usar recuentos totales para datos geográficos (por ejemplo, total de asesinatos) puede ser engañoso, ya que a menudo se correlaciona con el tamaño de la población. En su lugar, usar tasas (por ejemplo, tasa de asesinatos por cada 100.000 personas) normaliza por la población y proporciona una comparación más precisa.¹ Los mapas de perfil geográfico que son esencialmente solo mapas de población son una trampa común.¹

Representación de la Incertidumbre

Importancia: Incluir estimaciones de incertidumbre o barras de error es una práctica crítica en la visualización científica, ya que mejora la toma de decisiones y proporciona un reflejo más preciso de la comprensión científica.¹

Diagramas de Caja y Bigotes (Box and Whisker Plots): Estos representan visualmente la distribución, el error o los intervalos de confianza.

- **Limitaciones:** Los diagramas de caja pueden percibirse como una unidad única y pueden no mostrar toda la distribución. También pueden hacer que diferentes grupos parezcan casi idénticos cuando no lo son, y el tamaño de la caja podría asociarse engañosamente con la cantidad.¹
- **Alternativas:** Los diagramas de violín (Violin Plots), los gráficos de gradiente (Gradient Plots), los gráficos de frijoles (Bean Plots) y los gráficos de enjambre de

abejas (Beeswarm Plots) ofrecen vistas más completas de la distribución de datos.¹

Bandas de Error y Fancharts: Para los gráficos de líneas, estas son alternativas recomendadas a los bigotes tradicionales para representar la incertidumbre.¹

Las advertencias recurrentes contra los ejes truncados, las líneas base no cero para las barras y las métricas engañosas ¹ no son solo cuestiones de "buen diseño", sino de "honestidad". Estas técnicas pueden manipular la percepción de forma intencionada o no. Esto subraya la responsabilidad ética del visualizador de datos. Cada elección de diseño, desde la escala de los ejes hasta la selección de las métricas, tiene el potencial de distorsionar la realidad e influir en las conclusiones. Por lo tanto, la precisión visual no es solo una habilidad técnica, sino un imperativo ético, especialmente en campos como la elaboración de informes industriales o la comunicación científica, donde las decisiones se basan en estas visualizaciones.

El énfasis en el uso de tasas en lugar de totales ¹ y la importancia crítica de incluir estimaciones de incertidumbre ¹ señalan un cambio de simplemente mostrar "lo que sucedió" a "cuánto" y "cuán seguros estamos". Esto sugiere que la visualización de datos madura va más allá de la estadística descriptiva para incorporar aspectos inferenciales. Proporcionar contexto a través de métricas normalizadas y cuantificar la incertidumbre capacita a la audiencia para tomar decisiones más informadas, reconociendo la variabilidad inherente o las limitaciones en los datos. Esto es crucial para informes analíticos robustos y publicaciones científicas.

VI. Interactividad y Animación

La interactividad y la animación transforman las visualizaciones estáticas en herramientas dinámicas y exploratorias, mejorando la participación del usuario y la comprensión.

Interactividad: Ventajas y Conceptos Clave

Ventajas: La interactividad amplía los límites físicos de lo que se puede ver en un

espacio determinado, cubre una mayor variedad de análisis, facilita la manipulación de datos, amplifica el control del usuario y amplía el interés.¹ Transforma al observador en un usuario proactivo.¹

Conceptos Clave:

- **Evento:** Una acción del usuario (por ejemplo, pulsar una tecla, hacer clic con el ratón).¹
- **Control:** La reacción del sistema a un evento (por ejemplo, mostrar una ventana emergente, resaltar una opción).¹
- **Acción:** La operación realizada en la aplicación (por ejemplo, filtrar, resaltar información).¹
- **Data Linking (Vinculación de Datos):** Cuando se seleccionan datos en un gráfico, la misma posición/elemento se resalta en otros.¹
- **Slider (Deslizador):** Un componente para seleccionar un rango de valores.¹
- **Toggle (Conmutador):** Un interruptor para seleccionar opciones.¹
- **Utility (Utilidad):** Si el sistema hace lo que los usuarios esperan.¹
- **Efficiency (Eficiencia):** La rapidez con la que los usuarios pueden realizar las tareas.¹

Categorías de Interacción

- **Filtrado:** Permite a los usuarios especificar qué datos ver u ocultar, modificando el mensaje comunicado.¹ Ejemplos incluyen el uso de botones, ventanas emergentes, deslizadores o casillas de verificación.¹
- **Selección:** Resalta los valores de interés sin eliminar otros, a menudo modificando los atributos de color o reordenando los datos.¹ Ejemplos: resaltar premios Nobel, nombres de bebés o usar el "brushing" en histogramas.¹
- **Participación/Colaboración:** Permite a los usuarios contribuir con sus propios datos para personalizar la representación visual.¹ Ejemplos incluyen cuestionarios o dibujar en mapas.¹
- **Resumen:** Proporciona diferentes niveles de anotaciones de datos bajo demanda, permitiendo al usuario controlar la percepción de los datos.¹ Ejemplos: ventanas emergentes para información adicional.¹
- **Observación/Exploración:** Muestra múltiples vistas o diferentes niveles de detalle de forma dinámica, especialmente cuando no toda la información puede mostrarse simultáneamente.¹ Ejemplos: zoom/pan en mapas, desglose o

exploración lineal a través de secuencias narrativas.¹

Animación: Propósito, Ventajas y Consideraciones Clave

Propósito: La animación muestra datos con una dimensión temporal de forma dinámica, permitiendo apreciar patrones que no son visibles en imágenes estáticas.¹ Crea la ilusión de movimiento, generando un fuerte impacto visual.¹

Ventajas: Aumenta la dimensionalidad de los datos (tiempo), proporciona información diferente a los gráficos estáticos y ofrece calidad estética.¹ Puede guiar transiciones, enfatizar elementos y mostrar cambios a lo largo del tiempo.¹

Consideraciones Clave:

- **Duración entre fotogramas:** Demasiado lento puede ser aburrido, demasiado rápido puede llevar a la pérdida de detalles.¹
- **Potencial de distracción:** La animación puede distraer del mensaje.¹
- **Control del usuario:** Es crucial que el usuario tenga control para iniciar, detener y controlar la animación (por ejemplo, botones de reproducción/pausa, deslizadores para los fotogramas).¹

Tabla 4: Definiciones de Conceptos de Interactividad y Usabilidad

Concepto	Definición
Evento	Una acción iniciada por el usuario (ej. clic, pulsación de tecla).
Control	La reacción del sistema a un evento (ej. mostrar un pop-up, resaltar una opción).
Acción	La operación que se realiza en la aplicación (ej. filtrar, resaltar información).
Data Linking	La interconexión de datos entre múltiples gráficos, de modo que la selección en uno resalta elementos correspondientes en los demás.

Slider	Un componente de interfaz de usuario que permite seleccionar un valor o un rango de valores a lo largo de un continuo.
Toggle	Un control de interfaz de usuario que permite alternar entre dos estados (ej. activado/desactivado).
Data Usability	La calidad de los datos y la información en el contexto de la visualización de datos, abarcando fiabilidad, estabilidad y soporte a la toma de decisiones.
Data Reliability	La calidad de los procesos de datos que inspiran confianza en los datos utilizados, considerando los intervalos de error.

Las ventajas de la interactividad, como la expansión de los límites de lo visible, la mayor variedad de análisis y el control del usuario ¹, abordan directamente el problema de la sobrecarga de información en conjuntos de datos grandes y complejos. Esto permite una divulgación progresiva y una exploración personalizada en lugar de gráficos estáticos y densos. Esto implica que la interactividad no es solo una característica añadida, sino un principio de diseño fundamental para la visualización de datos moderna, especialmente en contextos dinámicos y de alta dimensionalidad. Cambia la carga de trabajo del visualizador, que antes tenía que digerir toda la información, a capacitar al usuario para que encuentre sus propios conocimientos, convirtiendo la visualización en una herramienta de descubrimiento y no solo de presentación.

La animación, si bien es elogiada por su "fuerte impacto visual" y su capacidad para mostrar la "dimensión temporal de forma dinámica" ¹, también conlleva riesgos significativos. Las advertencias sobre que puede ser "demasiado lenta (aburrida) o demasiado rápida (pérdida de detalles)" y que "puede distraer del mensaje" ¹, junto con el énfasis en el "control del usuario" ¹, revelan una doble filo. Esto sugiere que, si bien la animación puede ser muy atractiva y efectiva para mostrar cambios a lo largo del tiempo, también conlleva un riesgo considerable de sobrecarga cognitiva o interpretación errónea si no se implementa con cuidado. La capacidad del usuario para controlar el ritmo y el enfoque de la animación es fundamental para mitigar estos riesgos, transformando un elemento potencialmente distractor en una valiosa herramienta analítica. Esto subraya la importancia del diseño centrado en el usuario en las visualizaciones dinámicas.

VII. Usabilidad y Experiencia de Usuario (UX)

La usabilidad y la Experiencia de Usuario (UX) son fundamentales para evaluar si los sistemas de visualización de datos sirven eficazmente a sus usuarios previstos, especialmente en el caso de los sistemas interactivos.

Usabilidad: Definición y Componentes

La usabilidad es un atributo de calidad que evalúa la facilidad de uso de aplicaciones, sitios web, interfaces y sistemas por parte del usuario.¹ Incluye cinco componentes clave: Aprendizaje (Learnability), Eficiencia (Efficiency), Memorabilidad (Memorability), Errores (Errors) y Satisfacción (Satisfaction).¹ La usabilidad es una condición necesaria para la adopción por parte del usuario. Los usuarios abandonan los sistemas que son complicados, que fallan con frecuencia o que no son intuitivos. Está estrechamente relacionada con la productividad.¹

Utilidad vs. Usabilidad vs. Útil

- **Utilidad:** Se refiere a la funcionalidad, es decir, si el sistema hace lo que los usuarios esperan.¹
- **Usabilidad:** Se refiere a lo fácil y satisfactorio que es el uso del sistema.¹
- **Útil:** Es la combinación de Utilidad y Usabilidad.¹

Usabilidad de Datos: Fiabilidad, Estabilidad, Soporte a la Toma de Decisiones

La usabilidad de datos se refiere a la calidad de los datos y la información en el contexto de la Visualización de Datos.¹ Se asocia con principios como la Fiabilidad de los Datos (calidad del procesamiento de datos, confianza en los datos), la Estabilidad

de los Datos (minimizar el impacto negativo de transformaciones como la reducción de dimensionalidad) y el Soporte a la Toma de Decisiones (representación lo suficientemente clara para decisiones informadas).¹

Experiencia de Usuario (UX): Definición e Indicadores Clave

La UX estudia la experiencia general de un usuario al interactuar con un sistema.¹ Es una consecuencia del estado interno del usuario, las características del sistema y el contexto de interacción.¹ Sus indicadores clave abarcan la usabilidad, la interfaz de usuario, la experiencia de interacción, el diseño interactivo y la experiencia del cliente, incluyendo aspectos objetivos, estéticos y emocionales.¹ Los métodos de evaluación de la UX incluyen evaluaciones de expertos, estudios de laboratorio, estudios de campo, cuestionarios, métodos mixtos y medidas psicofisiológicas.¹

Modelos de UX: Atributos Pragmáticos vs. Hedónicos

Los modelos de UX se basan en la usabilidad y la emotividad, considerando elementos externos como el usuario, el contexto y los aspectos temporales.¹

- **Atributos Pragmáticos (Instrumentales):** Se refieren a la utilidad y la funcionalidad, medidos por aspectos como la facilidad de aprendizaje, la eficiencia y la efectividad. Se relacionan con la manipulación del sistema.¹ Para la visualización de datos, las cualidades instrumentales incluyen la utilidad, la completitud, la percepción, la confianza y la intuición.¹
- **Atributos Hedónicos (No Instrumentales):** Son aspectos no instrumentales como la estética, el comportamiento, la funcionalidad intuitiva, la estimulación (por ejemplo, descubrir sorpresas), la identificación y la evocación (recordar experiencias anteriores).¹ Para la visualización de datos, estos incluyen la estética y el atractivo.¹

Modelos específicos de UX incluyen:

- **Hassenzahl UX:** Distingue entre atributos pragmáticos y hedónicos. Utiliza el cuestionario Attrakdiff (28 criterios, 7 escalas).¹
- **User Experience Questionnaire (UEQ):** Evalúa productos interactivos con 26

indicadores en 6 categorías: Atractivo, Perspicuidad, Eficiencia, Fiabilidad, Estimulación y Novedad.¹

- **meCUE:** Una herramienta de evaluación modular basada en el modelo CUE, que define características instrumentales y no instrumentales, así como respuestas emocionales. Es flexible y adecuada para sistemas de visualización de datos.¹ Su cuestionario tiene 34 ítems en tres módulos: percepciones del producto (instrumentales/no instrumentales), emociones y consecuencias de la UX.¹

Cuestionarios Relevantes

- **SUS (System Usability Scale):** 10 preguntas, escala Likert de 5 puntos. Rápido, simple y fiable para la usabilidad general.¹
- **PSSUQ (Post-Study System Usability Questionnaire):** 16 preguntas, escala Likert de 7 puntos. Más complejo, cubre la Utilidad del Sistema, la Calidad de la Información y la Calidad de la Interfaz.¹
- **Attrakdiff:** Utilizado para el modelo Hassenzahl UX, se centra en las cualidades pragmáticas y hedónicas.¹
- **meCUE:** Recomendado como el más adecuado para sistemas de Visualización de Datos, ya que equilibra correctamente la usabilidad con la estética visual y es flexible.¹

Tabla 3: Clasificación de Preguntas de Usabilidad y UX

PREGUNTA	CATEGORÍA	TIPO DE TEST
1. I thought the System was easy to use	a) Instrumental / Usabilidad / Pragmáticas	SUS
2. Non-inclusive 1 2 3 4 5 6 7 Inclusive	b) No instrumentales / Hedónicos / Emocionales	Attrakdiff, Hedonic
3. I found the various functions in this System very well integrated	a) Instrumental / Usabilidad / Pragmáticas	SUS
4. Complicated 1 2 3 4 5 6 7 Simple	a) Instrumental / Usabilidad / Pragmáticas	Attrakdiff, Pragmatic

5. The product is creatively designed	b) No instrumentales / Hedónicos / Emocionales	meCUE, A1, Aesthetic
6. The product is stylish: Strongly Disagree 2 3 4 5 6 Strongly Agree	b) No instrumentales / Hedónicos / Emocionales	meCUE, A3, Aesthetic
7. It was easy to find the information I needed	a) Instrumental / Usabilidad / Pragmáticas	PSSUQ
8. This System has all the functions and capabilities I expect it to have	a) Instrumental / Usabilidad / Pragmáticas	PSSUQ
9. The design looks attractive	b) No instrumentales / Hedónicos / Emocionales	meCUE, A2, Aesthetic
10. Conservative 1 2 3 4 5 6 7 Innovative	b) No instrumentales / Hedónicos / Emocionales	Attrakdiff, Hedonic

El fuerte énfasis en que "la usabilidad es una condición necesaria para que los usuarios utilicen el sistema" y que los usuarios "abandonan el sistema si es complicado" ¹, junto con la definición de "útil" como la combinación de "utilidad" (funcionalidad) y "usabilidad" (facilidad de uso) ¹, implica que incluso una visualización técnicamente perfecta (alta utilidad) fracasará si no es usable o proporciona una mala experiencia de usuario. Esto subraya que la precisión técnica en el procesamiento de datos, la elección de algoritmos (por ejemplo, DR) y la codificación visual (por ejemplo, teoría del color, precisión) deben complementarse con un fuerte enfoque en la UX. El análisis más sofisticado resulta inútil si el usuario no puede comprenderlo o interactuar con él de forma eficaz. Esto resalta la importancia del diseño iterativo y las pruebas de usuario en el ciclo de vida de desarrollo de cualquier producto de visualización de datos, trascendiendo la mera ciencia de datos para abarcar la interacción persona-ordenador.

La observación de que la "calidad hedónica (emotividad) disminuye con el tiempo, al dejar de ser una novedad" ¹, en contraste con la calidad pragmática que "aumenta con el tiempo" ¹, tiene implicaciones significativas. Esto sugiere que los factores de "sorpresa" iniciales, como animaciones complejas o diseños novedosos, pueden atraer a los usuarios, pero no mantendrán el compromiso a largo plazo. Para un producto de visualización de datos sostenible, el diseño debe priorizar las cualidades pragmáticas (eficiencia, facilidad de aprendizaje, fiabilidad) que mejoran con el uso repetido, mientras que los elementos hedónicos podrían necesitar una actualización periódica o ser diseñados para evocar conexiones emocionales más profundas y

duraderas, en lugar de una novedad fugaz. Esto tiene implicaciones directas para los ciclos de vida del desarrollo de productos y la priorización de características en las herramientas de visualización de datos.

Obras citadas

1. examens-all.pdf