Charla 16: Técnicas de Pronóstico de Calidad del Aire: Regresiones Estadísticas y CART

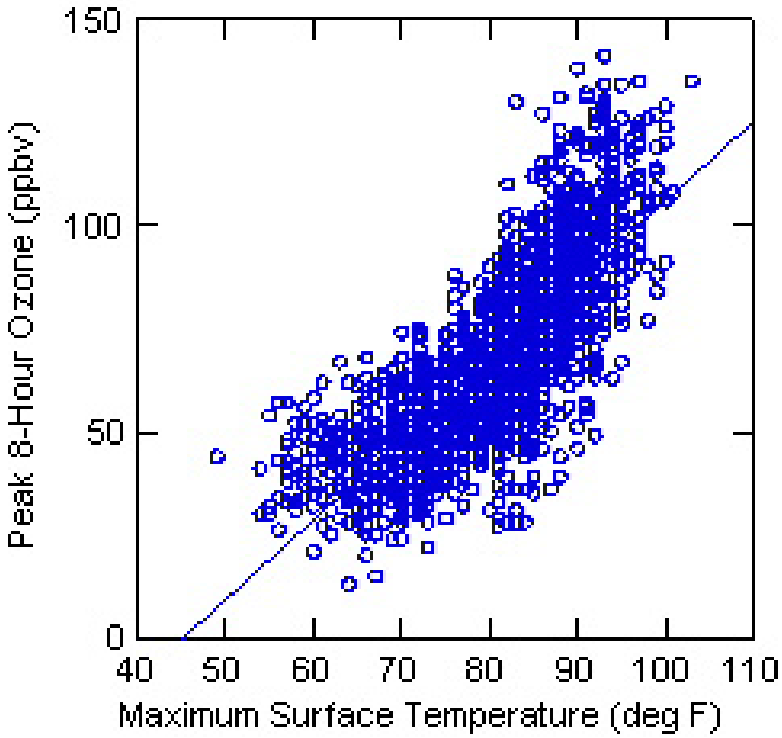
Taller Centroamericano de Pronosticado de la Calidad del Aire

San José, Costa Rica

17-21 de Octubre del 2011

Ecuaciones de Regresión Estadística

- Las ecuaciones de regresión estadísticas usan **variables predictoras** para calcular la concentración esperada de un contaminante, como el O_3 o $PM_{2.5}$
- La forma común de las regresiones lineales es y = mx + b
- Las variables predictoras pueden ser meteorológicas o de calidad del aire que afectan las concentraciones del contaminante
- Ejemplo: la temperatura máxima (T_{max}) es un buen predictor para el O_3 máximo.



$[O_3] = 1.92*T_{max} - 86.8$

$r = 0.77 \quad r^2 = 0.59$

Ecuaciones de Regresión Estadística

- Más predictores se pueden adicionar a una "Regresión paso a paso": $y = m_1x_1 + m_2x_2 + m_3x_3 + \ldots\ldots m_nx_n + b$
- Cada variable predictora (x_n) tiene su propio "peso" (m_n) y la combinación normalmente lleva a obtener una mejor precisión en el pronóstico.
- La combinación de las variables depende del área de pronóstico (i.e., los predictores para San José, Costa Rica no serán los mismos que para San Salvador, El Salvador)
- Software estadístico se utiliza para determinar la ecuación de regresión

Variables Predictoras Comunes para $PM_{2.5}$

Variable	Utilidad	Condición de alta $PM_{2.5}$
500-mb altura	Indicador del patrón climático de escala sinóptica	Alta
Intensidad del viento superificia	Asociado a la dispersión y dilución de contaminantes	Baja
Dirección del viento superficial	Asociado con el transporte de contaminantes	-
Gradiente de presión	Causa viento/ventilacion	Baja
Cocentración pico de $PM_{2.5}$ del día anterior	Persistencia, arrastre	Alta
850-mb temperatura	Sustituto del mezclado vertical	Alta
Precipitación	Associated with clean-out	Ninguna o ligeramente
Humedad Relativa	Affects secondary reactions	Alta
Vacaciones	Additional emissions	-
Día de la Semana	Emissions differences	-

Ejemplo de la ecuación de regresión

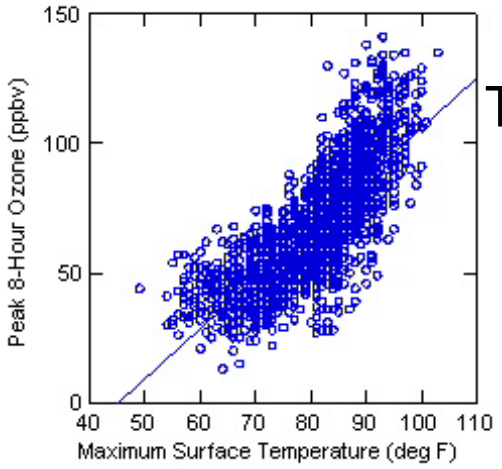
Ecuación de regresión de ozono de Columbus, Ohio

$$8hr\ O_3 = \exp(2.421 + 0.024*T_{max} + 0.003*T_{range} - 0.006*WS_{1to6} + 0.007*00Z_{V925} - 0.004*RH_{Sfc00} - 0.002*00Z_{WS500})$$

Variable	Descripción
T_{max}	Temperatura máxima en ºF
T_{range}	Intervalo de temperatura durante el día
WS_{1to6}	Velocidad media del viento de 1 p.m. a 6 p.m. en nudos
$00Z_{V925}$	Componente V del viento a 925-mb a las 00Z
RH_{Sfc00}	Humedad relatica en superficie a las 00Z
$00Z_{WS500}$	Velocidad del viento a 500 mb a las 00Z

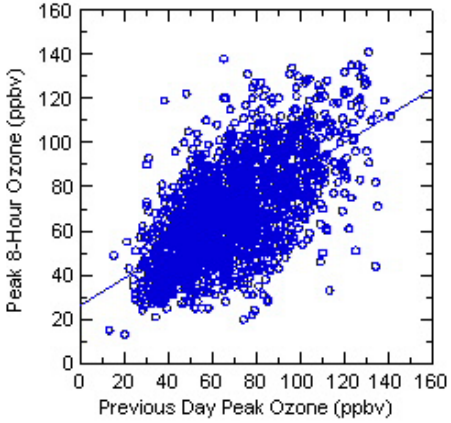
Ecuaciones de Regresión

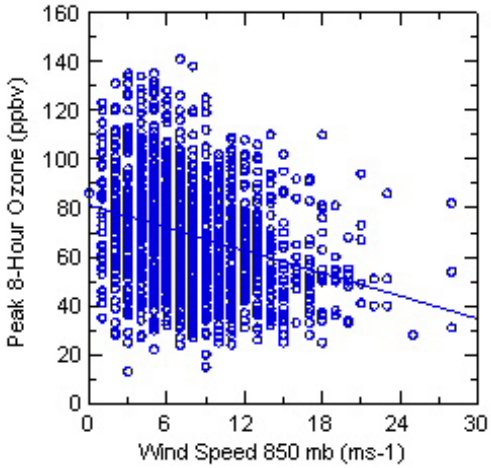
- Las diferentes variables predictoras no son ponderadas igualmente ya que unas son más importantes que otras.
- Es fundamental identificar los predictores más fuertes y trabajar más en conseguir esas predicciones correctas



T_{max} vs. O_3

Día anteiror O_3 vs. O_3



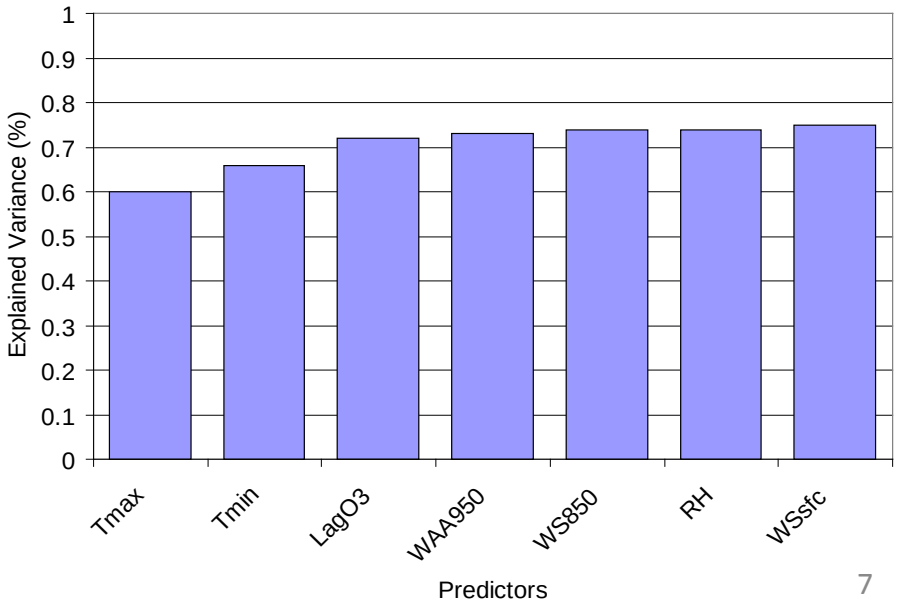


Vel. de viento vs. O_3

Ecuaciones de Regresión

- En un caso de ejemplo, la mayor parte de la varianza del O_3 es explicada por T_{max} (60%), los predictores adicionales añadirán ~ 15%
- En general, 75% de la varianza observada en el O_3 es explicada por el modelo de previsión.
- Nuestra tarea como pronosticadores es completar el 25% adicional con otras herramientas.

Varianza explicada acumulada



Desarrollo de Ecuaciones de Regresión

- Determinar los procesos importantes que influyen en las concentraciones de contaminantes
- Seleccione las variables que representan los procesos importantes que influyen en las concentraciones de contaminantes
- Crear un conjunto de datos multi-anual de las variables seleccionadas
  - Elegir los últimos años que son representativos del perfil actual de emisiones
  - Reserve un subconjunto de datos para una evaluación independiente, pero garantice que representan todas las condiciones.
  - Asegure que las variables son pronosticadas
- Utilice el software estadístico para calcular los coeficientes y las constantes de regresión de la ecuación
- Realizar una evaluación independiente del modelo de regresión

Desarrollo de Ecuaciones de Regresión

- Emplear el log natural de la concentración de contaminantes como predictor puede mejorar el desempeño
- No “sobre ajuste” el modelo mediante muchas variables.
  - Un “sobre ajuste” podrá disminuir la exactitud del pronóstico
  - Un número razonable de variables es de 5 a 10
- Variables únicas deben de usarse ara evitar redundancia o co-linearidad.
- La estratificación de los datos puede mejorar el desempeño de la regresión
  - Estaciones
  - Día laboral vs. fin de semana

Variables Predictoras

- La variables predictoras puede ser variables observadas (e.g., Conc. de $PM_{2.5}$ de ayer) y pronosticadas (e.g., Temperatura máxima de mañana).
- Asegúrese que las variables predictoras son fácilmente obtenibles de fuentes confiables o pueden pronosticasrse facilmente.
- Las variables de la predictoras deben capturar los fenómenos importantes que afectan a las concentraciones de contaminantes en la región
- Considere la posibilidad de incertidumbre en las mediciones, especialmente en las PM.

Selección de Variables Predictoras

- Comience con el mayor número de variables predictoras 50 a 100
- Use técnicas estadíticas para identificar las variables más importantes:
  - Utilizar el análisis de clúster para dividir los datos en subconjuntos similares y diferentes; use variables únicas (i.e., diferentes) para evitar redundancia.
  - Utilizar el análisis de correlación para evaluar la relación entre el predictando (es decir, los niveles de contaminantes) y las diversas variables predictor
  - Use regresión paso a paso en un software estadístico (como SAS, Statgraphics, Systat, STATA) para seleccionar las variables más importantes y generar la mejor ecuación de regresión
  - La Selección humana es otra manera de seleccionar las variables de predicción más importante

Creación de base de datos de variables predictoras

- Determinar los datos a usar:
  - ¿Qué tipo de datos y variables son necesarios?
  - ¿Qué sitios de monitoreo son representativos?
  - ¿Qué tipo de redes de monitoreo de calidad del aire a usar?(P.e. monitoreo continuo, pasivo)
  - ¿Qué tipo de datos meteorológicos estan disponibles?(superficie, en altura, satélire, etc)
  - ¿Cuántos años de datos están disponibles?

Adquisición de datos históricos

- Datos de monitoreo horario
- Métricas máximas diarias de contaminantes, como promedio de 24 horas $PM_{2.5}$ o PM_{10}
- Datos meteorológicos horarios
- Orientación con el modelo meteorológico y de calidad del aire
- Diagramas de superficie y en altura del tiempo
- Trayectorias de HYSPLIT

Control de calidad de la base de datos de las variables predictoras

- Comprobación de valores atípicos
  - Observar los valores máximo y mínimo para cada campo ;¿don razonables?
  - Compruebe la rapidez de cambio entre los registros de cada extremo
- Marcas de tiempo
  - ¿Todos los datos coinciden correctamente con el tiempo?
  - Gráficos de series de tiempo pueden ayudar a identificar los problemas de desplazamiento de UTC a LT
- Falta de datos
  - ¿Se utiliza el mismo identificador para cada campo? I.e., –999
- Unidades
  - ¿Las unidades son consistentes entre las diferentes bases de datos? p.e., m/s o knots para intensidad de viento
- Códigos de Validación
  - ¿Son los códigos de validación consistentes entre las bases de datos?
  - ¿Los códigos de validación coinciden con los valores de datos? P.e., ¿Son los datos faltantes marcados como -999?

Regresiones estadísticas: fortalezas

- Una técnica bien documentada que se utiliza ampliamente en varias disciplinas
- El software estadístico está disponible ampliamente
- Técnica de pronóstico que reduce los posibles prejuicios derivados de la subjetividad humana
- Puede sopesar adecuadamente las relaciones entre variables que son difíciles de cuantificar subjetivamente
- Se puede utilizar en combinación con otros métodos de pronóstico, o puede usarse como el método principal

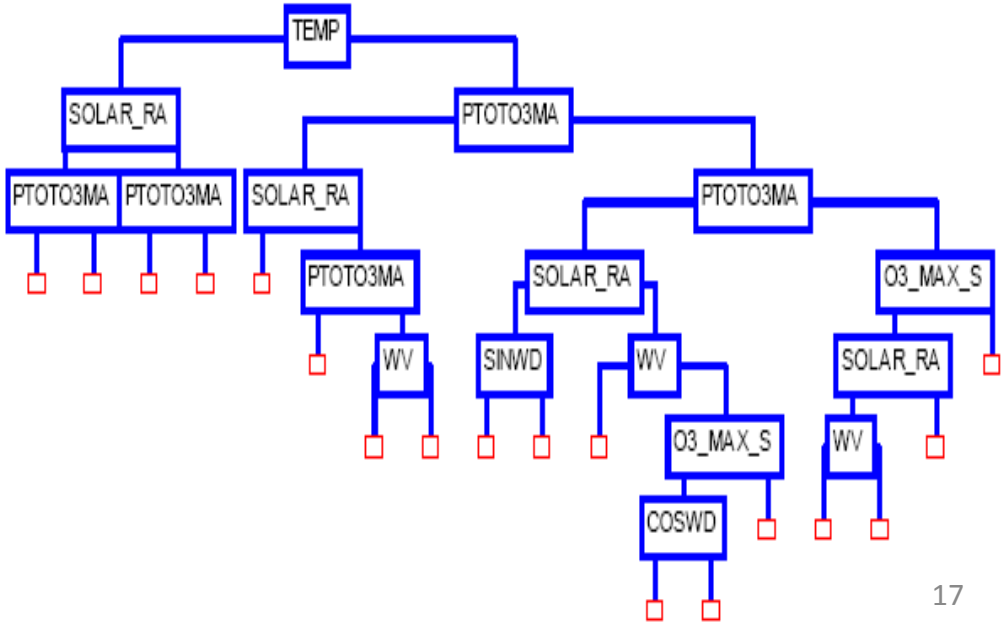
Regresiones estadísticas: limitaciones

- Las ecuaciones de regresión requiere de una cierta experiencia y esfuerzo para desarrollarlas
- Las ec. De regresión tienden a predecir mejor la media que los extremos (p.e., las concentraciones mayores del contaminante) de la distribución:
  - Subestima las concentraciones altas
  - Sobre estima las concentraciones bajas
- Las ecuaciones de regresión requieren actualizaciones periódicas de fuentes de emisión y los cambios de uso de suelo
- Las ecuaciones de regresión requieren de 3 a 5 años de datos medidos en la región, que incluyan los eventos de contaminación de aire.

Clasificación y Arbol de Regresion (CART)

- CART es un procedimiento estadístico diseñado para clasificar los datos en grupos diferentes.
- CART permite al pronosticador desarrollar un árbol de decisión para predecir las conc. De contaminantes basado en las variables de pronóstico (normalmente meteorológicas) que se correlacionan bien con el contaminante

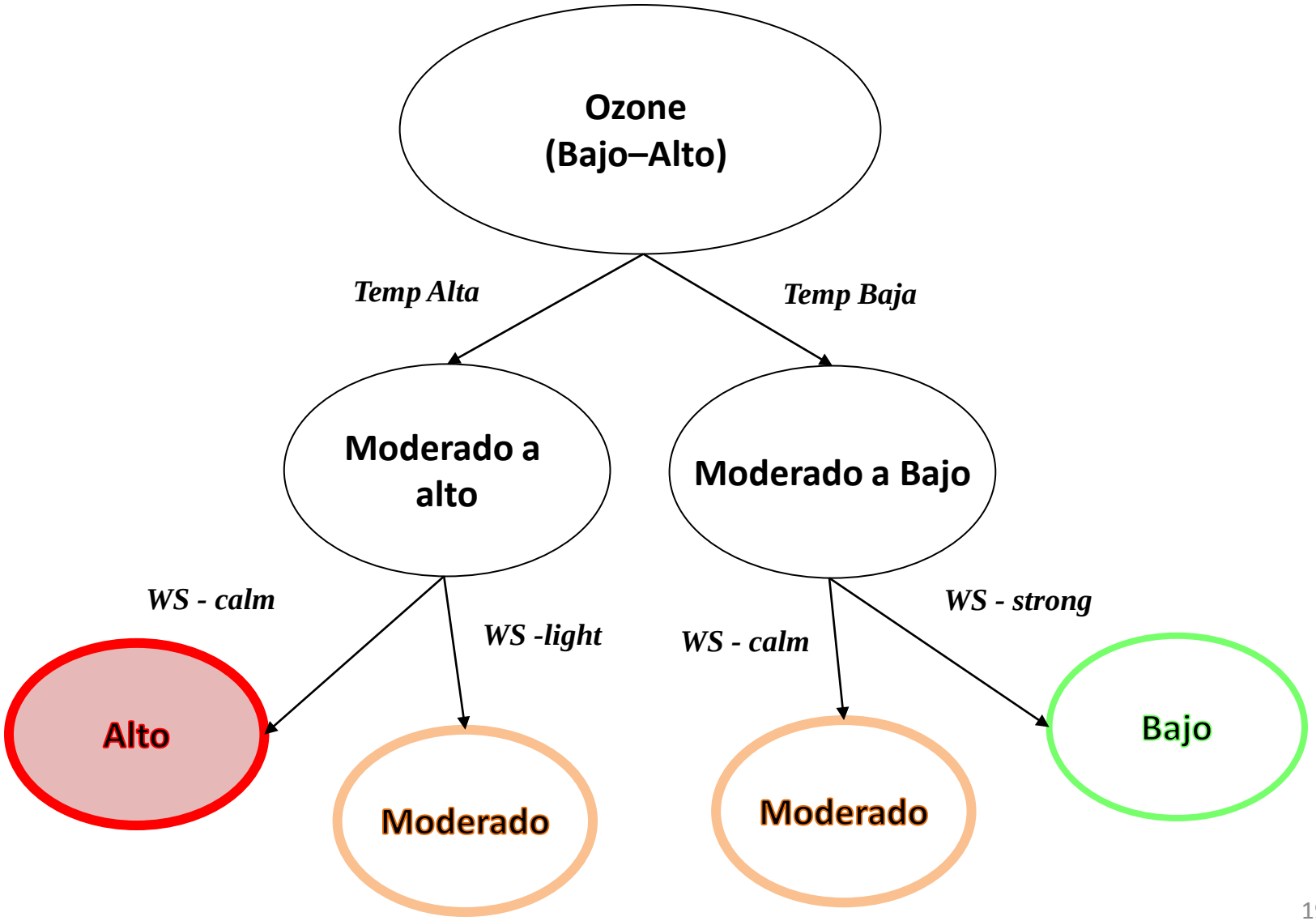
Un ejemplo de CART para la predicción de ozono máxima en el área metropolitana de Atenas



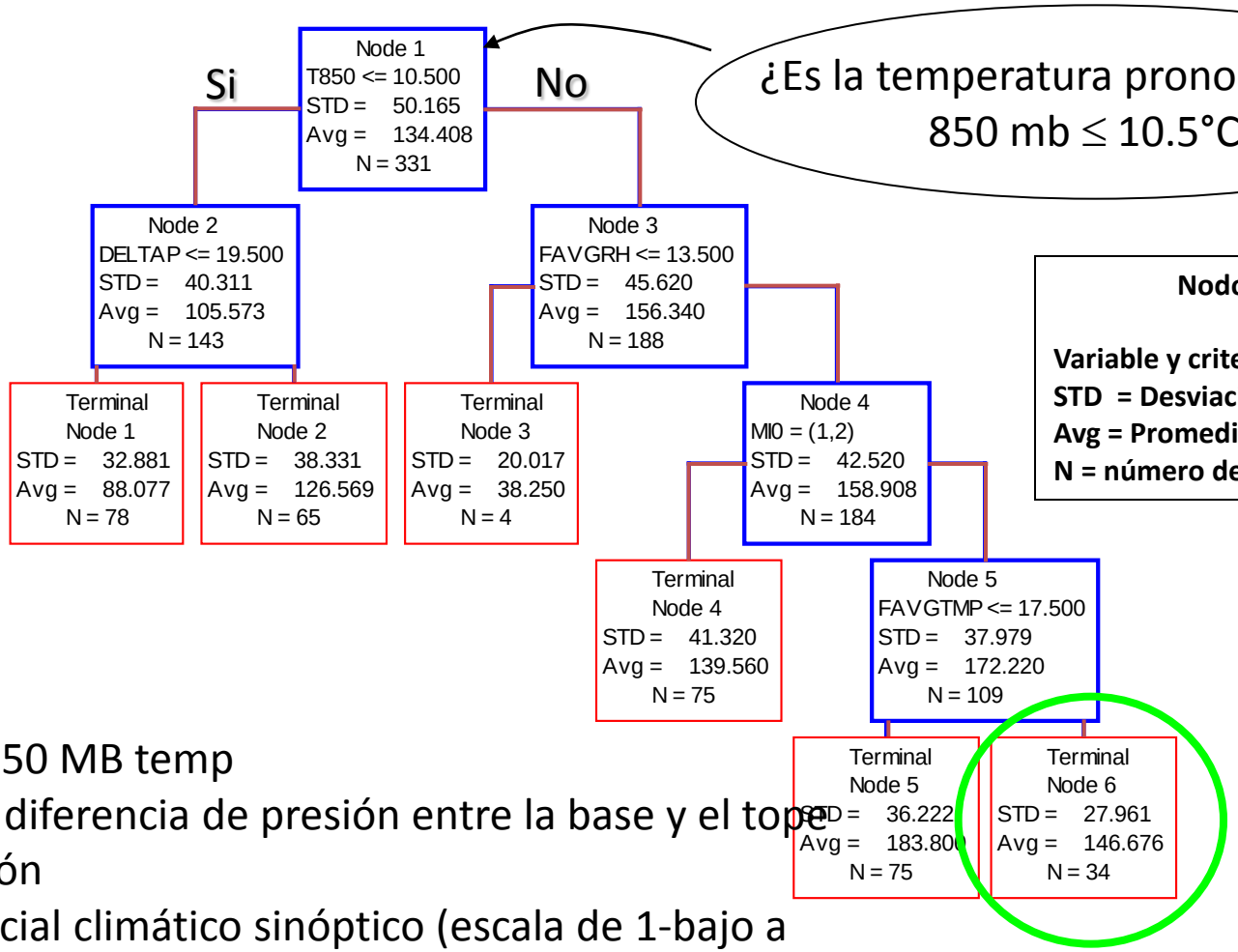
Pronósticos con CART

- Aplicando CART se inicia con la primera división y se determina en cual de los dos grupos pertenece el dato, basado en el valor de corte de esa variable.
- Continuar con CART de esa manera hasta que se llegue al nodo final
- La concentración media mostrada al final del nodo es la concentración pronosticada.
- Las diferencias ligeras en los valores de las variables predictoras pueden producir cambios significativos en los niveles de contaminantes pronosticados cuando sue valor es cercano al del umbral

Ejemplo CART para Ozono



Ejemplo PM_{10} CART para Santiago, Chile



Variables:

T850 - 12Z 850 MB temp

DELTAP – La diferencia de presión entre la base y el tope de la inversión

MI0 – Potencial climático sinóptico (escala de 1-bajo a 5-alto).

FAVGTMP – temperatura 24-hour prom en La Paz

FAVGRH – RH promedio 24-hour en La Paz

Cassmassi, 1999

Desarrollo de CART

- Determine los procesos importantes que influyen en la contaminación
- Seleccione las variables que representa adecuadamente los procesos importantes.
- Cree un conjunto de datos multi-anual de las variables seleccionadas
  - Selecciones los años recientes que son representativos del perfil actual de emisión.
  - Reserve un subconjunto de datos para una evaluación independiente, asegurando que representan todas las condiciones.
  - Asegúrese de que las variables son pronosticadas
- Emplee un software estadístico para crear el árbol de decisión
- Evalúe el árbol de decisión empleando el conjunto de datos independiente

CART - fortalezas

- Es un método de clasificación
- CART no requiere de seleccionar variables de antemano
- CART puede manejar fácilmente valores atípicos
- CART es flexible y se puede ajustar en el tiempo.
- CART no considera ninguna hipótesis y es computacionalmente rápido.

CART: limitaciones

- Requiere de experiencia y esfuerzo en el desarrollo
- Pequeños cambios en las variables predichas puede inducir grandes cambios en las concentraciones pronosticadas
- CART no puede predecir las concentraciones de contaminantes durante períodos donde se tiene patrones de emisión inusuales debido a días feriados u otros eventos
- Los criterios y enfoques estadísticos en CART requieren de actualizaciones debido a las emisiones y cambios de uso de suelo.