Universidad de Buenos Aires
Facultad De Ingeniería
Año 2019 - 1er Cuatrimestre

Germán Pérez Fogwill

Plan de tesis

1. Título y datos

Título:

Estimación de los parámetros en la formación de partículas atmosféricas usando métodos de procesamiento de señales y aprendizaje automático.

Alumno:

Nombre:	Germán Andrés Pérez Fogwill
DNI:	34270254
Padrón:	91709
Correo:	gfogwill@fi.uba.ar
Teléfono:	11-64845320

Tutora:

Nombre:	Patricia Pelle
DNI:	14715997
Correo:	pelle.patricia@gmail.com
Teléfono:	11-68529711

Co-tutora:

Nombre:	Eija Maria Asmi
DNI:	95870036
Correo:	eija.asmi@fmi.fi
Teléfono:	11-60384457

2. Objeto y área de la tesis

Los aerosoles son el conjunto de partículas microscópicas, sólidas o líquidas, que se encuentran en suspensión en un gas. En la atmósfera, estos afectan el balance energético de la Tierra, es decir, la diferencia entre la energía aportada por el sol y la energía que emite la Tierra. Sin embargo, la influencia que tienen los aerosoles en este balance energético esta asociada con grandes incertidumbres debido al desconocimiento de las fuentes, las concentraciones, y sus propiedades. Uno de los procesos más importantes asociado con la concentración de aerosoles es la formación de nuevas partículas (FNP). Este proceso, que consta de dos partes: *i*) la formación inicial de las partículas, llamado nucleación y *ii*) su posterior crecimiento, es la principal fuente de partículas presentes en la atmósfera terrestre. Los eventos de FNP típicamente son analizados por investigadores de forma manual a partir de la distribución de tamaños de partículas, lo que consume mucho tiempo y puede generar inconsistencias entre los resultados obtenidos debido a la subjetividad del análisis.

Para obtener resultados más confiables y consistentes la tendencia actual se enfoca en automatizar la extracción de parámetros de los eventos de FNP (tasa de crecimiento y de formación). En este trabajo se analizará la posibilidad de modelar la FNP utilizando Modelos Ocultos de Markov (ya sea solos o en combinación con otras herramientas de aprendizaje

automático) para poder identificar y cuantificar los eventos de FNP de forma automática.

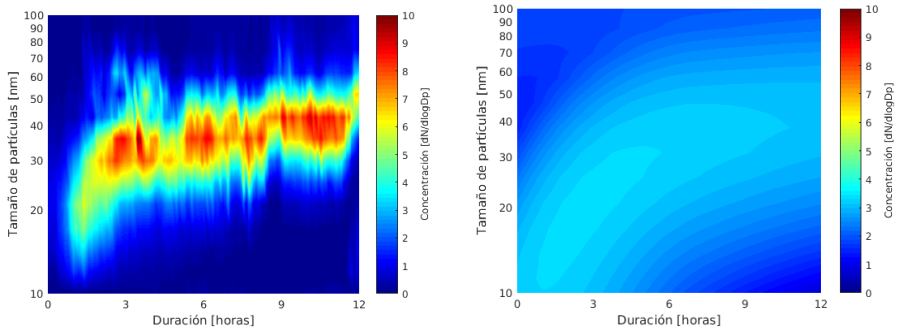
3. Introducción y antecedentes

Los aerosoles se encuentran distribuidos por toda la atmósfera terrestre e influyen en la calidad de vida de distintas maneras. En ambientes urbanos, los aerosoles afectan la salud de las personas a través de su inhalación [1, 2]. En la tropósfera, los aerosoles contribuyen al estado del clima y son un elemento esencial en el estudio del cambio climático [3–6]. Comprender sus efectos requiere información detallada de la manera en que los aerosoles ingresan a la atmósfera, cómo se transforman y cómo son removidos de la misma.

Las partículas son introducidas en la atmósfera ya sea por emisión directa (fuentes primarias) o por nucleación de gases de baja volatilidad (fuentes secundarias). El proceso que involucra la nucleación y el posterior crecimiento generalmente se llama "Formación de nuevas partículas" (NPF; *New Particle Formation*). Este proceso inicia con la formación de una partícula estable (nucleación) que después crece a través de la condensación [7]. Estudiando los parámetros de la formación de partículas (tasa de formación y tasa de crecimiento principalmente) se puede obtener información sobre los vapores que intervienen en el proceso, e inferir el tipo de nucleación, es decir si es homogénea o heterogénea. Experimentos realizados en laboratorio permiten explicar los diferentes tipos de nucleación que probablemente son responsables de la formación de nuevas partículas atmosféricas; sin embargo, aún se cuentan con pocos estudios de campo que permitan conocer con certeza en que medida estos procesos se dan en diferentes regiones de la atmósfera [8].

Se considera que hay un evento de FNP si se observa durante varias horas la formación y crecimiento de nuevos aerosoles comenzando desde el modo de nucleación (partículas con diámetro $< 25\,\text{nm}$). Estos eventos son identificados a partir de la evolución temporal de la distribución de tamaño de las partículas. La forma más sencilla y directa para detectar y cuantificar la FNP es utilizando métodos visuales a partir de las mediciones de distribución de tamaño de las partículas.

Mediante el uso de un medidor de partículas de movilidad diferencial (DMPS; *Differential mobility particle sizer*) es posible obtener la evolución temporal de la distribución de tamaños de las partículas.



(a) (b)

(a) Medición de un evento de formación de partículas. (b) Simulación de un evento de formación de partículas.

Típicamente este análisis requiere entre uno y tres investigadores para analizar períodos de formación y crecimiento de nuevas partículas en los datos de distribución de tamaño. A pesar que esos métodos son sencillos y directos de aplicar presentan varios inconvenientes. Primero, el análisis manual es una tarea intensiva debido a la falta de automatización. Segundo, los resultados obtenidos son subjetivos debido a que diferentes investigadores pueden interpretar de forma distinta la misma serie de datos.

La posibilidad de automatizar el análisis de la FNP ya fue investigada con anterioridad. Los dos métodos mas recientes y de mayor relevancia utilizan redes neuronales artificiales (ANN; *Artificial Neural Network*) para decidir si se produjo la FNP o no. El primero, utiliza redes neuronales convolucionales (CNN; *Convolutional Neural Network*) para detectar los eventos directamente a partir de los gráficos de la distribución de tamaño de las partículas [9]. El segundo, utiliza redes neuronales Bayesianas (BNN; *Bayesian Neural Network*) para decidir si se produjo la FNP o no [10]. Por el momento, ambos métodos sólo permiten identificar los momentos cuando se produce la FNP, quedando pendiente la parametrización del evento. El porcentaje de aciertos obtenidos por estos métodos es del 80 % aproximadamente (comparando este método contra la clasificación realizada por personas).

Estructuras y algoritmos basados en modelos ocultos de Markov (HMM; *Hidden Markov Model*) proveen una base matemática para construir sistemas de reconocimiento. Los HMM son usados para describir fenómenos aleatorios que están gobernados por mecanismos ocultos que son inaccesibles. En particular, los HMM se usan para describir sistemas que son observados en tiempo discreto, donde las observaciones son inducidas por un proceso subyacente ("oculto") que es desconocido. El aspecto estadístico de los HMM fue considerado por primera vez por Baum y Petrie [11]. Los HMM's se utilizaron anteriormente para el reconocimiento de formas temporales con éxito (como reconocimiento del habla, de escritura manual, de gestos).

Los sistemas híbridos HMM-ANN para el reconocimiento del habla fueron introducidos anteriormente por Bourlard [12]. Este tipo de sistemas combina la capacidad de los HMM's de modelar series temporales con la capacidad de clasificación de las ANN's. Trabajos recientes en redes profundas hacen uso de este tipo de sistemas híbridos para determinar las configuraciones óptimas de acuerdo a las cantidades de datos y recursos computacionales disponibles [13–15].

En este trabajo se evaluara la posibilidad de aplicar modelos de aprendizaje automático (ML; *Machine Learning*) para identificar y cuantificar la FNP. En particular se intentaran utilizar los HMM's, ya sea en su versión pura, o en su forma híbrida.

4. Desarrollo previsto de la tesis

4.1. Teoría, enfoque y métodos a utilizar

El primer paso en el desarrollo de la tesis sera la recolección y análisis de los datos con los que se entrenaran y evaluaran los modelos estadísticos. Se intentaran utilizaran datos de un DMPS instalado en la estación antártica Marambio. Se cuentan con cinco años de datos de distribución de tamaño de partículas atmosféricas. En comparación con las publicaciones antes mencionadas [9,10], se dispone de aproximadamente un 60 % menos de datos que los utilizados por ellos. La principal desventaja de las teorías propuestas en estos trabajos es que el espacio muestral disponible para crear la arquitectura de una red neuronal es inmensa debido a la cantidad de bloques de construcción y combinaciones posibles de los mismos [16]. Consecuentemente, el índice de aciertos del modelo va a depender fuertemente de la selección de los bloques [17], la cual puede ser realizada por los investigadores o de forma automática. A diferencia de los trabajos anteriores, se cambiará el enfoque intentando hacer un máximo aprovechamiento de los datos explotando la naturaleza temporal de los eventos de FNP a través de los HMM's.

En caso que las mediciones reales no sean suficientes o no sean lo suficientemente buenas (debido al ruido de medición, o a *gaps* en los datos), se dispone con un modelo desarrollado por la Universidad de Helsinki (UHEL) que permite realizar simulaciones de eventos de FNP. Este modelo simula la química y la dinámica de los aerosoles. También se cuenta con la asesoría de expertos en el tema de la UHEL y del Instituto Meteorológico Finlandés (FMI; *Finnish Meteorological Institute*). Estas simulaciones pueden ser utilizadas para dos fines: *i*) tener más datos para entrenar los sistemas estadísticos; y *ii*) testear si los resultados obtenidos son correctos.

El segundo paso consistirá en plantear un modelo de detección de FNP y utilizar los datos (ya sean mediciones reales o simulaciones) para entrenar dicho modelo. En esta etapa también se probará si el modelo planteado detecta de forma satisfactoria (es decir, con un índice de aciertos del mismo orden que en los trabajos antes mencionados) o si es necesario cambiar el enfoque del modelo propuesto. Como primera aproximación se intentará utilizar HMM's puros, y en caso de no obtener los resultados esperados se evaluaran distintos modelos híbridos comparando los resultados para el mismo conjunto de datos de entrenamiento.

4.2. Estudios conexos

El alumno ya tiene aprobadas las siguientes materias relacionadas con los temas relacionados con la tesis:

- Probabilidad y estadística (61.09)
- Señales y sistemas (66.74)
- Procesos estocásticos (66.75)
- Procesamiento de señales I (66.38)
- Procesamiento de señales II (66.39)
- Procesamiento del habla (66.46)
- Procesamiento de imágenes (66.47)

- Teoría de la detección y estimación (66.51)

De forma simultanea con el desarrollo de la tesis, el alumno va a cursar las siguientes asignaturas durante el año 2019:

- Introducción a proyectos (66.12) - 1^{er} *Cuatrimestre 2019*
- Procesos Markovianos para aprendizaje automático - Instituto de cálculo (FCEN-UBA) - 1^{er} *Cuatrimestre 2019*
- Redes neuronales (66.63) - 2^{do} *Cuatrimestre 2019*

Desde el día 22 de julio hasta el día 2 de agosto el alumno asistirá al taller "*Atmospheric Aerosols: Properties, Measurements, Modeling, and Effects on Climate and Health*" a llevarse a cabo en la Universidad de São Paulo (`https://sites.google.com/view/spsas-aerosols/spsas`).

4.3. Alcances de la tesis

En esta investigación se pretende desarrollar una herramienta que permita eliminar la subjetividad al momento de cuantificar los eventos de FNP. El objetivo propuesto es lograr un sistema capaz de: *i*) detectar la FNP con igual o mayor índice de aciertos que en las publicaciones mencionadas; y *ii*) extraer los parámetros que caracterizan el evento detectado.

El planteo del modelo de detección no se limitara solamente a los HMM's tradicionales, sino que además se podrá investigar la posibilidad de utilizar HMM's no paramétricos y redes neuronales convolucionales. También se planteara la posibilidad de combinar los distintos métodos.

Durante la etapa de entrenamiento de los sistemas se compararan las mediciones reales contra las simulaciones disponibles desde un punto de vista estadístico. Se podrá también plantear un nuevo modelo que permita simular la FNP, pero a diferencia del modelo provisto (el cual tiene su escénica en la física y química) se le dará un enfoque estadístico (modelo generativo).

En la fase final del trabajo se compararán los parámetros obtenidos de forma automática contra los parámetros obtenidos por los investigadores.

Referencias

[1] A. Ibald Mulli, H. E. Wichmann, W. Kreyling, and A. Peters, "Epidemiological Evidence on Health Effects of Ultrafine Particles," *Journal of Aerosol Medicine*, vol. 15, no. 2, pp. 189–201, 2002.

[2] D. M. Stieb, S. Judek, and R. T. Burnett, "Meta analysis of time-series studies of air pollution and mortality: Effects of gases and particles and the influence of cause of death, age, and season," *Journal of the Air and Waste Management Association*, vol. 52, no. 4, pp. 470–484, 2002.

[3] K. E. Trenberth, J. T. Fasullo, and J. Kiehl, "Earth'S Global Energy Budget," *Bulletin o the American Meteorological Society*, vol. 90, pp. 311–323, 2009.

[4] P. A. Stott, S. F. B. Tett, G. S. Jones, M. R. Allen, J. F. B. Mitchell, and J. Jenkins, "External Control of 20th Century Temperature by Natural and Anthropogenic Forcings," *American Association for the Advancement of Science Stable*, vol. 290, no. 5499, pp. 2133–2137, 2016.

[5] V. Ramanathan, P. J. Crutzen, J. T. Kiehl, and D. Rosenfeld, "Atmosphere: Aerosols, climate, and the hydrological cycle," *Science*, vol. 294, no. 5549, pp. 2119–2124, 2001.

[6] S. Menon, A. D. D. Genio, D. Koch, and G. Tselioudis, "GCM Simulations of the Aerosol Indirect Effect: Sensitivity to Cloud Parameterization and Aerosol Burden," *Journal of the Atmospheric Sciences*, vol. 59, no. 3, pp. 692–713, 2002.

[7] M. Kulmala, H. Vehkamäki, T. Petäjä, M. Dal Maso, A. Lauri, V. M. Kerminen, W. Birmili, and P. H. McMurry, "Formation and growth rates of ultrafine atmospheric particles: A review of observations," *Journal of Aerosol Science*, vol. 35, no. 2, pp. 143–176, 2004.

[8] T. Jokinen, M. Sipilä, J. Kontkanen, V. Vakkari, P. Tisler, E. M. Duplissy, H. Junninen, J. Kangasluoma, H. E. Manninen, T. Petäjä, M. Kulmala, D. R. Worsnop, J. Kirkby, A. Virkkula, and V. M. Kerminen, "Ion-induced sulfuric acid–ammonia nucleation drives particle formation in coastal Antarctica," *Science Advances*, vol. 4, no. 11, pp. 1–7, 2018.

[9] J. Joutsensaari, M. Ozon, T. Nieminen, S. Mikkonen, T. Lähivaara, S. Decesari, M. C. Facchini, A. Laaksonen, and K. E. Lehtinen, "Identification of new particle formation events with deep learning," *Atmospheric Chemistry and Physics*, vol. 18, no. 13, pp. 9597–9615, 2018.

[10] M. A. Zaidan, V. Haapasilta, R. Relan, H. Junninen, P. P. Aalto, M. Kulmala, L. Laurson, and A. S. Foster, "Predicting atmospheric particle formation days by Bayesian classification of the time series features," *Tellus, Series B: Chemical and Physical Meteorology*, vol. 70, no. 1, 2018.

[11] L. E. Baum and T. Petrie, "Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to The Annals of Mathematical Statistics. ® www.jstor.org," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[12] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition.* 1994.

[13] J. Paisley and L. Carin, "Hidden markov models with stick-breaking priors," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3905–3917, 2009.

[14] C. A. McGrory and D. M. Titterington, "Variational Bayesian analysis for hidden Markov models," *Australian and New Zealand Journal of Statistics*, vol. 51, no. 2, pp. 227–244, 2009.

[15] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. M. A. Patwary, Prabhat, and R. P. Adams, "Scalable Bayesian Optimization Using Deep Neural Networks," 2015.

[16] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L. J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive Neural Architecture Search," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11205 LNCS, pp. 19–35, 2018.

[17] M. Ghassemi, L. Lehman, J. Snoek, and S. Nemati, "Global optimization approaches for parameter tuning in biomedical signal processing: A focus on multi-scale entropy," *Computing in Cardiology 2014*, no. Mimic, pp. 993–996, 2014.