We introduce a new method for forecasting emergency call arrival rates that combines integer-valued time series models with a dynamic latent factor structure. Covariate information is captured via simple constraints on the factor loadings. We directly model the count-valued arrivals per hour, rather than using an artificial assumption of normality. This is crucial for the emergency medical service context, in which the volume of calls may be very low. Smoothing splines are used in estimating the factor levels and loadings to improve long-term forecasts. We impose time series structure at the hourly level, rather than at the daily level, capturing the fine-scale dependence in addition to the long-term structure. Our analysis considers all emergency priority calls received by Toronto EMS between January 2007 and December 2008 for which an ambulance was dispatched. Empirical results demonstrate significantly reduced error in forecasting call arrival volume. To quantify the impact of reduced forecast errors we design a queueing model simulation that approximates the dynamics of an ambulance system. The results show better performance as the forecasting method improves. This notion of quantifying the operational impact of improved statistical procedures may be of independent interest. KEYWORDS: Ambulance planning; Dynamic factor model; Non-homogeneous Poisson process; Integer-valued time series; Smoothing splines.